Docketed

COPY

JUL 9 1999

UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF WASHINGTON
AT SEATTLE

****

THE NORTHWEST LABORERS-EMPLOYERS HEALTH & SECURITY TRUST FUND, and its TRUSTEES; THE CARPENTERS HEALTH AND SECURITY TRUST OF WESTERN WASHINGTON and its TRUSTEES; THE WASHINGTON-IDAHO CARPENTERS-EMPLOYERS HEALTH & WELFARE TRUST, and its TRUSTEES; THE WASHINGTON-IDAHO OPERATING ENGINEERS HEALTH AND WELFARE TRUST FUND, and its TRUSTEES,

Plaintiffs

vs. No. C97-849WD

PHILIP MORRIS, INCORPORATED; R.J. REYNOLDS TOBACCO CO.; BROWN & WILLIAMSON TOBACCO CORP.; B.A.T INDUSTRIES P.L.C.; BRITISH AMERICAN TOBACCO COMPANY, LTD.; LORILLARD TOBACCO CO.; THE AMERICAN TOBACCO CO., INC.; LIGGETT & MYERS, INC.; UNITED STATES TOBACCO CO.; THE TOBACCO INSTITUTE, INC.; THE COUNCIL FOR TOBACCO RESEARCH U.S.A., INC.; SMOKELESS TOBACCO COUNCIL, INC.; HILL AND KNOWLTON, INC.; and UNKNOWN CORPORATIONS, A-Z,

Defendants

****

VOLUME: I
PAGES: 1 - 114

DEPOSITION of DONALD B. RUBIN, Ph.D

JULY 2, 1999

produced in part by the University of California, San Francisco Library
http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Digitized by the Internet Archive
http://www.industrydocuments.ucsf.edu/docs/hygl0001

52299 4586

DEPOSITION of DONALD B. RUBIN, Ph.D., a witness called on behalf of the Plaintiffs pursuant to the Federal Rules of Civil Procedure, before Judith McGovern Williams, Certified Shorthand Reporter No. 130993, Registered Professional Reporter, Certified Realtime Reporter, and Notary Public in and for the Commonwealth of Massachusetts, at the offices of Goodwin, Procter & Hoar, One Exchange Place, Conference Room 22A, Boston, Massachusetts 02109, on Friday, July 2, 1999, commencing at 3:10 p.m.

APPEARANCES:

STRITMATTER, KESSLER, WHELAN & WITHEY
[By Michael E. Withey, Esquire]
1200 Market Place Tower
2025 First Avenue
Seattle, Washington 98121
206-448-1777
On behalf of the Plaintiffs

Produced by LTI, Inc. in HUMPHREY

52299 4587

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

APPEARANCES (Continued):

HELLER, EHRMAN, WHITE & McAULIFFE
[By John W. Phillips, Esquire]
6100 Columbia Center
701 Fifth Avenue
Seattle, Washington 98104-7098
206-389-6126
All on behalf of the Defendant
Philip Morris, Incorporated

JONES, DAY, REAVIS & POGUE
[By Peter J. Biersteker, Esquire and
Brian C. Castello, Esquire]
51 Louisiana Avenue, N.W.
Washington, D. C. 20001-2113
202-879-3939
Both on behalf of the Defendant
R. J. Reynolds Tobacco Co.

produced by R.J.R.T.C. in HUMPHREY

52299 4588

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

INDEX

Witness	Page
DONALD B. RUBIN, Ph.D.	
Direct Examination by Mr. Withey	5

EXHIBITS

Number		Page
1	Document headed Northwest Laborers Litigation Supplemental Report of Professor Donald B. Rubin	78
2	Second Supplemental Report of Donald B. Rubin - Northwest Laborers	78
3	Article entitled Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders	88

Produced by IF IC in EMPHREY

52299 4589

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

PROCEEDINGS

- - -

DONALD B. RUBIN, first having been duly sworn, testified as follows in answer to direct examination by MR. WITHEY:

- - -

Q. Would you state your name and give your business address for the record, please?

A. Donald B. Rubin. My primary affiliation is the Department of Statistics at Harvard University.

Q. I will need the address.

A. One Oxford Street, Cambridge, Mass. 02138.

Q. Dr. Rubin, you have submitted I guess it has been two reports I think in this case. Is that your understanding? Or is it three?

A. I believe it is three.

Q. Okay. And we have -- one was on -- did you happen to bring those with you?

A. I have copies somewhere with me, but not on the desk here, the table here.

Produced by Tobacco Industry Publishing
52299 4590
http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. September 11, 1998; February 23 1999; and May 26, 1999?
A. That sounds right.
Q. And you understood that the purpose of submitting these reports was to provide us with your opinions and the basis of your opinions in this case? Correct?
A. Correct.
Q. And you understood that there was some time deadlines that were required in submitting these reports?
A. Correct.
Q. And you understood that in submitting the work that you had done in preparing this report that they -- that that work was to have been completed, with the exception of preparing for testimony, by the times the reports were to be completed? Correct?
A. If I understand the question, yes.
Q. All right.
A. That there were things that had deadlines, and I was -- I was to make those deadlines in the reports.
Q. Right. And as I understand it, one of the

Produced by R.J. Reynolds in litigation

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

areas of work that you have undertaken is to conduct what might for better term be called a multiply or multiple imputation analysis of data in the National Medical Expenditure Survey?

A. Basically that is right, although it is not a multiple imputation analysis per se. It is creating a multiply-imputed version of NMES that can support valid analyses.

Q. And you have been -- you were testifying --

MR. WITHEY: Strike that.

Q. You were deposed in both this case as well as in the Ohio ironworkers' litigation by myself? Correct?

A. Correct.

Q. And the last time we met for deposition, I believe it was sometime in January in the ironworkers' case, you testified about the process by which the multiple imputation work was being undertaken? Correct?

A. Yes. We discussed it I think relatively briefly, but the issue certainly arose.

Produced by RJ Reynolds Tobacco Company

52299 4592

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. All right. I am not going to ask you -- the reason I wanted to make sure we understand each other is I am not going to go back and ask all the same questions, because you testified then, I assume honestly, about your work in performing this analysis?

A. Yes. I did.

Q. And you understood that in that case, the Ohio case again, there were deadlines that were required to be fulfilled in order to submit your work to us, you know, in that case? Correct?

A. Correct.

Q. And it was your intention, was it not, to have completed the multiple imputation -- what is the word you would be comfortable with?

A. NMES.

Q. Multiple imputation of NMES --

A. Yes.

Q. -- by the time of your deposition in January in the ironworkers' case; correct?

A. I was hoping to have that done. Correct.

52299 4593

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. And as I sit here today, that multiple imputation of NMES has not been provided to us. Is that correct?

A. That's correct.

Q. And could you describe for me all efforts you have undertaken to complete that work from the time we last deposed you, and I will ask you about this, in January to the present?

A. I'm not sure what exactly you mean by all the efforts, but I will describe in general what all the efforts have been.

Q. Fair enough.

A. They have involved further work on my designing what that multiple imputation process should look like using two resources. One resource was a former student of mine. He is at the University of Michigan. And the other resource is somebody at Wecker Associates. Because it's a -- it's a major effort to try to multiply impute NMES, and, in fact, it is more complicated than I had first imagined, and it is more complicated than

produced by RJR in HUMPHREY

52299 4594

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

the other major multiple imputation projects in which I have been involved.

Q. Okay. I appreciate the general description. Now I would like to specify -- be a little more specific. First of all, what is the name of the person at the University of Michigan that has been working on this project?

A. His name is Dr. Raghunathan.

Q. Could you spell that?

A. R A G H U N A T H A N.

Q. Two names?

A. No. One name.

Q. Oh, okay. And when was he first contacted to perform this work?

A. Let's see. Do I really know? It was before the January deposition that you are referring to. Maybe a -- maybe a month before. Maybe a little before that. I am uncertain. I could figure it out if I went back and looked at records, but -- but that's the closest I can come just sitting here right now.

Q. What records are you referring to that you

Produced in LITIGATION by RJRTC

52299 4595

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Available from www.industrydocuments.ucsf.edu/docs/hygl0001

have of work performed on this project?
A. Well, I would have some records of phone calls to him; days and hours that I spent working on the project.
Q. Any correspondence?
A. Written correspondence or --
Q. e-mails?
A. We did do some e-mailing. I don't think any of it has been saved, because most of the discussions took place by telephone.
Q. Okay.
A. I mean I think the kind of e-mails which were sent were, "Have you done that yet," sort of e-mails that don't convey any information. It would just be a message that he could come and pick up without having to return a phone message.
Q. How about data exchanged? Is that reflected in any document or electronic mail?
A. Well, there were some intermediate sort of diagnostic outputs that were looked at along the way, but none of that was saved.
Q. Was there any other category of records,

Produced by R.J. Reynolds Tobacco Company in the Litigation Document Depository

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by www.industrydocuments.ucsf.edu/docs/hygl0001

either electronic or actual hard copy, that would reflect the work that you have done along with the work of the doctor in Michigan and William Wecker Associates?

A. I don't believe so.

Q. Okay. And where are those records located?

A. You mean the nonexistent ones? I don't --

Q. I thought you made records of phone calls.

A. Oh, phone calls? Oh. There are -- Raghu -- his name is Raghunathan -- collected records of his hours, which he would transmit to me, and I have records of my own hours that I had spent on it.

Q. And where are they located?

A. They are on a calendar -- mine are on a calendar, and his presumably are on a calendar that he keeps.

Q. They would be a record of the time spent and the day on which they were spent?

A. Yes.

Q. Could you please collect all of those and provide those to Mr. Biersteker, so that I can have them marked as Exhibit 1 for this

Produced by the UCSF Industry Documents Library

http://legacy.library.ucsf.edu/tid/oqq07a00/pdf
Support for www.industrydocuments.ucsf.edu/docs/hygl0001

deposition? Would you be willing to do that?

A. Okay. I could. I mean I can go back and try to -- and try to do that.

Q. Thank you.

A. Okay.

MR. BIERSTEKER: We will take that request under advisement.

MR. WITHEY: Fine. Let me know if you have decided not to do that, Peter.

MR. BIERSTEKER: I will let you know.

MR. WITHEY: So we can raise it as soon as possible.

MR. BIERSTEKER: All right.

MR. WITHEY: In fact, just for the record, our motions in limine I think are due on July 29th? Is that right, John?

MR. PHILLIPS: 29th.

MR. WITHEY: July 29th. Thank you. So we would want to resolve any dispute about our discovery of those records by a week before then so that we

Produced by the UCSF Industry Documents Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

could obtain them and attach them, if necessary.

MR. BIERSTEKER: All right.

BY MR. WITHEY:

Q. Doctor, could you describe the work that Raghu did for you on this project?

A. Sure. It will have to be slightly technical. I will try to make it as nontechnical as possible.

Q. Thank you.

A. The multiple imputation of NMES was actually divided into -- I divided it into three phases. And there is a first phase, which is the computationally far more expensive and complicated phase, and that's the phase that I tried to -- I have been working with Raghunathan, and the second and third phase are the phases that were primarily allocated to -- I allocated to the Wecker Associates.

The first phase was -- took the basic NMES data and all of the variables, we call them covariates, which are the nonexpense data, and made a summary for

produced by RJRTC
in the TUPLEY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

each person of the expense data, which arrived in a form which we -- created in the form of a table for each person of expense, expenses for each type and ICD 9 code, and to take that data set, which of course generated approximately 500 variables for each person, and tried to multiply impute that. What multiply impute means is that you just do the same thing repeatedly using random draws to represent the uncertainty about what the right value to impute is.

The procedure being used is or was -- sorry. Let me start that sentence again.

The fundamental idea of this kind of multiple imputation is based on ideas that are sometimes called Markov Chain Monte Carlo methods. They are a way of simulating very complex distributions. The ideas go back more than half a century but have become quite popular in statistics and related research areas in the last decade as computational equipment

Produced by IRJRTC in HOUSTON

52299 4600

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

has made it possible.

It is a way of multiply -- it is a way of drawing simulated values in an iterative way. What I mean by that is you don't do it in one pass. You have to keep cycling around until the procedure stabilizes. Even one run on relatively sophisticated computing equipment today can take a day or two or three, which means that the process of debugging computer programs can be -- take a very long time.

Because NMES has many variables that we want to impute, all these detailed medical expenditure information by ICD 9 code, for example, this turns out to be a very demanding process.

Q. Have you completed your answer?

A. If -- if that answers adequate for your purposes, then I have.

If you want me to expand upon it, I can.

Q. Well, I thought you were going to describe which -- the three phases. I'm not sure

Produced by the IRTC in HUMPHREY

52299 4601

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the National Library of Medicine. http://www.industrydocuments.ucsf.edu/docs/hygl0001

what you just described fit into -- was that the first phase, or is that all three?
A. That was the first phase.
Q. Okay.
A. I believe your question was what I asked Raghunathan to do. That is what I asked him to do is the first phase. That is all I have described.
Q. So you did how many different computer runs that took one to three days did you run?
A. In trying to debug the thing, it must have been dozens. Maybe -- starting and figuring out problems and mistakes and rerunning things, maybe more than dozens.
Q. So each time that -- was it Raghu that did the runs then as opposed to yourself?
A. Raghunathan did all of the actual computer runs.
Q. What kind of computer does he have?
A. He has, I think, a couple of them of the very powerful new PCs. Pentium. I don't know which ones.

produced by RJRC in HUMPHREY

52299 4602
http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. When did he start work on phase one?

A. I believe back before the January deposition which you referred to.

Q. So this would have been in December of 1998?

A. I am guessing, but again I could try to recover that with a phone call or --

Q. Okay. And prior to that time, you had worked with Dr. Schafer at Penn. State? Correct?

MR. BIERSTEKER: I object to the form of the question.

A. Yes. Let me clarify that. I have worked with Professor Schafer on other multiple imputation projects and on other papers, because he was my Ph.D. student at Harvard. I never worked with him for the multiple imputation of NMES at all. I never worked with him on anything with regard to this particular case.

Q. Did you ask him to do that?

A. No. I did not. I gave his name to Peter Biersteker to contact, because Joe Schafer and I had worked on two major multiple

Produced by RJRTC in HUMPHREY

52299 4603

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

imputation projects together for the federal government, whose surveys are like this one.

We did EDHANES, E D H A N E S, and FARS, F A R S, together several years before, and so since I was very busy, I thought that these were projects that he might be -- that these were projects -- this project might be one that he would want to take on by himself, since he had quite a bit of experience working on such projects.

Q. And to your knowledge, did he agree to take on that project?

A. That is my understanding.

Q. And did he generate any output from that project that you are aware of?

A. Not that I'm aware of. I don't think he did. I did speak to him about it towards the end of his involvement, and he was just completely tied up with other things. He was in the process of his last year before potential tenure promotion, and he was very much involved in trying to get

52299 4604

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

papers published and accepted in journals to secure his future at Penn. State.

Q. Was it after that time when it became known that Dr. Schafer could not complete work on this project that you then asked Dr. Raghu -- how do you pronounce it?

A. Raghunathan.

Q. -- Raghunathan to work with you on doing it?

A. Yes. I -- I asked him, and I -- I spoke -- at that time I decided I would try to take charge of the project and try to figure out what resources were needed to actually do the multiple imputation and get into understanding what NMES was really -- the features of NMES that made it a difficult problem and how to decide a multiple imputation scheme that would work.

Q. What were the features of NMES that made it a difficult problem?

A. Well, one is the richness of the data. Generally there are lots of variables. In particular, the fact that there are a lot

produced by RJRTC
in the UCSF

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Digitized by the Internet Archive http://www.industrydocuments.ucsf.edu/docs/hygl0001

of expense variableS, medical expense variables that are critical for this litigation, that have to be considered at a fairly detailed level; in other words, not just total medical expense, but medical expense broken down by ICD 9 codes, because the Surgeon General says these certain ICD 9 codes have different relative risks associated with them and, therefore, carving them up in any kind of calculation for smoking-attributable expenditures you have to be able to carve out different amounts for different ICD 9 codes, and, therefore, you can't just deal with total medical expenditures. You have to deal with them at the level of ICD 9 codes.

So that means that you create many variables, many more variables than you would if you were just concerned with, for example, total expense.

Q. Any other? Was there significant missing data?

A. Yes. There was significant missing data

produced by RJRTC in tuvaplay

52299 4606

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Buttons adapted from www.industrydocuments.ucsf.edu/docs/hygl0001

but not unheard of by any means by federal database standards. It was the significant missing data in the sense of far too significant to handle by ad hoc methods.

Q. Any other attributables of the NMES data set that you claim made your work more complicated or complex or demanding?

A. Well, also there was the desire to handle missing information, missing data, on smoking characteristics, characteristics of smoking habits for people, because it was such an important variable for this case, and the fact that the way the questions were asked in NMES created a relatively complex structure on variables related to smoking.

Q. How many of the variables related to smoking did you analyze?

A. It is hard to answer that directly, because they -- because the data have kind of a nested structure. As I remember, that people were first asked, for example, whether they were -- had they ever smoked

Produced by R. J. C. Reynolds Tobacco Company in the Minnesota Tobacco Trial

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stroke http://www.industrydocuments.ucsf.edu/docs/hygl0001

cigarettes, and so if they said they never did, then they didn't pursue asking questions about when is the last time you smoked a cigarette, because it makes no sense.

Then if they did, then I think they asked them whether you were a current smoker or a former smoker, and then depending on that branch, how many cigarettes you smoke now versus how many cigarettes you smoked before and when you started and when you quit.

So it wasn't just a simple set of questions that were asked of every person, but there was this branching structure to it, which is more complicated to deal with, more complicated to model than a simple how old are you, how many cigarettes do you smoke, what is your mother's age when you were born, those simple, you know, somewhere everybody gets a question. That created complications for us as well.

Q. You testified there were 500 variables for

produced by RJRTC in LITIGATION

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

each person?
A. Approximately.
Q. Do you know how many variables there were for that related to smoking, smoking status, habits?
A. Well, let's see. Approximately five.
Q. That is because how many questions were asked on the survey, approximately?
A. It doesn't have this one-to-one correspondence --
Q. Right.
A. -- because the fact it has this nesting structure. For some people it asked one question and no more. For some it asked more.

One of the issues is how to code that information in such a way that it was most -- would be most reliable for future analyses. One of our objectives was to try to include in all the variables in NMES for which we were doing multiple imputation or using to do multiple imputation each of the variables that any of the plaintiffs used in any of the

Produced by R.J.R.C.
Tobacco Industry Documents

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

cases, in the smoking cases, as well as the variables that the defense used so we would have a complete data set that could be used to support analyses of any of the types that were -- had been used in these cases.

Q. Well, did you have access to the actual data from NMES?

A. I personally did not, but I believe both Raghunathan and the Wecker team did.

Q. Where did you get that?

A. Where did they get it? I don't know. I think -- well, I can say -- well, I know a little bit.

Q. Tell me what you know.

A. Okay. At one point, Raghunathan had the NMES data set that was used by Harrison in Oklahoma. Harrison did an analysis for the plaintiffs in the Oklahoma lawsuit, and in fact, that was one of the reasons why we were delayed. He was using that data set, and Harrison had done certain things to that data set -- I am not saying necessarily they are wrong or right -- but

produced by FRTC in HUMPHREY

52299 4610

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

they had done certain things to that data set that were not right for our purposes.
Q. When were --
A. And --
Q. Excuse me. When was that obtained, first obtained? Do you know?
A. Raghunathan must have obtained that at the time of the Oklahoma -- when the Oklahoma lawsuit was going on, when it was ongoing.
Q. Was Raghunathan involved as a defense statistician in any litigation prior to the Northwest laborers or ironworkers case?
A. I don't think he has ever been involved as a statistician except I have been using him, if that's what you mean.
Q. Yes.
A. I have been using him to do some analyses, primarily these propensity score analyses and now these multiple imputation work --
Q. So as I understand it then --
A. -- for --
Q. -- the Harrison data set was provided pursuant to discovery or whatever in the

produced by RJRTC in FUMPHREY

52299 4611

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Oklahoma Attorney General litigation? Correct?

A. I'm not sure, but presumably yes.

Q. All right. And Dr. Raghunathan had that data set in conjunction with the Oklahoma case?

A. Correct.

Q. Okay.

A. Correct.

Q. So, in other words, it wasn't like you had to get it, the Harrison data set, and send it to him for the first time as the result of this case; correct?

A. Correct. Oh, absolutely correct.

Q. Do you recall when that data set, the year or month that that data set was provided to Dr. Raghunathan?

A. No. I do not.

Q. It would have been in 1997 or 1998? Correct?

A. I think that has got to be right.

Q. Something like that?

A. Yes.

Q. So, in other words, had the defendants

Produced in Discovery

52299 4612

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Additional text www.industrydocuments.ucsf.edu/docs/hygl0001

asked you shortly thereafter to oversee the project of multiple imputation of NMES back when Dr. Raghunathan received the Harrison data set, you could have begun work then; correct?

A. Could have begun work at that time?

Q. Yes.

A. Although it turns out that that data set was not the right one to use in any case.

Q. Was there another data set?

A. I think -- I think that the data set that is now being used came from Wecker Associates, and I think that is the real NMES. What I mean by the real NMES is is it is not as filtered through some other data analysts' hands.

Q. When did Wecker get that?

A. I have no idea.

Q. Was it in conjunction with the tobacco litigation?

A. I have no idea.

Q. When did you find out that Wecker had this data set?

A. Well, this is a slight reconstruction, but

Produced by the Industry

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Powered by the Industry Documents Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

I think it is accurate. When we started
finding out that there were problems with
the multiple imputation of NMES that
Raghunathan was doing based on
communications between the Wecker folks
and us and inconsistencies that were
arising. We started realizing that we
weren't dealing with the same data set.
Q. And at that point in time, Wecker had the
data set that let's call it the Wecker
data set --
A. Okay.
Q. -- of NMES, and Raghunathan had the
Harrison data set of NMES?
A. Right. That's my understanding.
Q. And they were both doing some work on it,
and they realize there was something
different?
A. Yes.
Q. I am using some lay terms but.
A. That's fine.
Q. Okay.
A. What you are saying is basically right.
That's when we decided that, hey, there is

produced by RJRTC
in HUMPHREY
52299 4614

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

a real problem here.

Q. So assuming the Wecker data set was available as of day one, whatever date that is, and I don't mean to suggest when it was, you could have begun -- he could have begun work on this particular project as of that date?

A. Well, he -- he could have had I started to think about how to do it and the right way to do it.

Q. All right.

A. Yes.

Q. You don't know exactly when Wecker got his data set?

A. I do not.

Q. Do you understand -- who in Wecker & Associates have you been working with, particularly as it relates to phase two and phase three?

A. Primarily Gary Harvey.

Q. Spell it.

A. H A R V E Y.

Q. What does Mr. Harvey do? What is his background?

Confidential Subject to Protective Order

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. His background would be, I think, in applied -- he was in the Air Force, I think, and did applied mathematics, and he has quite a bit of knowledge of statistics as well, but I don't remember his degrees. In fact I don't even know if I ever have ever seen a CV for him.

Q. Well, other than he was in applied mathematics in the Air Force, then you don't know what his training was?

A. Well, he -- he did -- he did a lot of computer support work and statistical support work at Wecker Associates for quite a few years.

Q. Has he ever done a multiple imputation analysis before?

A. I don't think so.

Q. How about Dr. Raghunathan has?

A. Yes.

Q. Now have you completed your description of phase one?

A. Unless you want more detail.

Q. I may, but just your general description of phase one?

Produced by E.R.T.
in HOUSTON

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. I mean -- yes.
Q. Can you tell me about phase two, please?
A. Okay. Phase two took the output from phase one, which had parameter values in it, and actually imputed individual expense amounts for the events for which there was missing expense amounts.
I'm not sure I got the tenses right there and the singular and plural, but.
So the first phase produced, after all of this iteration, is designed to produce after all of this iteration certain which might be called parameter values that can be used to, as input to the second stage. The second stage takes those input parameters and for each event with missing expenditures attaches to that an imputed expense amount which is randomly drawn from a certain distribution.
Q. Is it within a range?
A. Pardon?
Q. Is it within a range?

produced by RJRTC in HUMPHREY

52299 4617

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

A. Yes. It is --
Q. With boundaries?
A. Well, it is -- it is imputed under log normal specification. Log of the data are normal. And it is subject to editing constraints that the values being imputed can't be larger or smaller than anything that is observed in the data.
Q. Okay. Who wrote the program for that application?
A. I designed it, and -- this is phase two -- and Gary Harvey supervised someone or maybe more than one person at Wecker Associates to actually write the code.
Q. And when did you design the program for phase two?
MR. BIERSTEKER: I object to the form.
A. My guess is December.
Q. December '98?
A. Probably early December.
Q. 1998?
A. Yes. 1998, I believe.
Q. And you were first asked to do that, to

52299 4618

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Tobacco Industry Documents ucsf.edu/docs/hygl0001

work on this, multiple imputation of NMES -- refresh my recollection. Was it around November, I think, 1998 or before then?

MR. BIERSTEKER: Objection. Asked and answered.

A. I'm not sure exactly when it was, but it followed the realization that Joe Schafer could not do it.

Q. Okay. Well, I think you answered that already.

Did you design any program for phase one?

A. I designed part of it. I guess I could add that phase one was utilizing a substantial programming effort that had already been implemented by Raghunathan at the Institute for Survey Research at the University of Michigan. So it wasn't all new code. It was taking that code that existed and had been used in various projects to implement the phase one.

Q. Well, do you know how many statisticians or people doing applied mathematics has

Produced in HUMPHREY

52299 4619

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Support provided by the UCSF Library
http://www.industrydocuments.ucsf.edu/docs/hygl0001

done a multiple imputation analysis in this country?

MR. BIERSTEKER: I object to the form.

A. It depends -- well, let me tell you why I am pausing.

There are really two levels, at least two levels you might be working on in order to do a multiple imputation. At the level we are doing it, we are doing it with a major data set. Major, I mean a large number of people, a large number of units, and a large number of variables.

Q. Right.

A. And for these, very often it requires a team of people to get involved in doing it. I have been involved in some of those, and I haven't been involved in others. There have probably been, let me guess, maybe five projects that are major in that sense. So in each of those projects, you will find more than one person involved.

Then there are a variety of what

produced by FTC in PRIVILEGE

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

might be called more minor data sets, maybe hundreds of units and dozens of variables, that people have done in published articles, and there may be dozens of people, maybe hundreds of people who have done those.

Q. Okay. We are talking about the former here rather than the latter; right?

A. Okay. If that's what you want to talk about, that's what we will talk about.

Q. In other words, of the kind done in this case, this is a major data set the way you have described it?

A. Absolutely.

Q. There have been five major --

A. Approximately.

Q. -- projects of that kind done that you are aware of? Correct?

A. There may be more.

Q. All right.

A. I would have to sit and actually think of which ones they were.

Q. And who all worked on those, if you know?

A. Well, --

Produced by RJRTC
Unpublished in Library

52299 4621
http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. And what were they?
A. -- one of the first that comes to mind I was slightly involved in maybe 10 years ago that was done at the Federal Reserve system by an Art Kennickal, K E N N I C K A L, I believe. And it is the -- what is it -- I think it is the Survey of Consumer Finances.
Q. Where is Art Kennickal?
A. He is at the Federal Reserve. I think he is still there.
Q. Okay.
A. And that was -- and that used the same kind of idea as is being done here, the Markov Chain Monte Carlo of the particular type that has been done here.
Q. Okay.
A. There is another effort -- I will do this in the order that comes to mind.
Q. I don't need --
A. At Erasmus University in the Netherlands, a guy named Jaap Brand, medical database.
Q. Spell it, please.
A. J A A P B R A N D at Erasmus University in

Produced by RJRTC
in EVIDENCE

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

Rotterdam. I think that was a fairly major application.

Q. On medical database?

A. Medical database, yes. There were other people involved in that, just as there were with Art Kennickal whose names I can try to remember. No?

Q. No.

A. There was EDHANES, a long project we did for the National Center of Health Statistics. I was involved at the beginning. Mainly I directed it. It also involved Joe Schafer and Ron Little; Trina Ezzatti-Rice, who is at NCHS; Meena Khari, who is at NCHS; what is his name, Johnson, I can't remember his first name, Wayne Johnson, who is at NCHS.

Q. Okay. That is three.

A. Okay. The Fatal Accident Reporting System, FARS, which is for the Department of Transportation, National Highway Traffic Safety Administration, which was done primarily by Joe Schafer with my designing aspects of it. He also used a

Produced by R.J. Reynolds in UNITED STATES litigation

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the University of California, San Francisco Library
http://www.industrydocuments.ucsf.edu/docs/hygl0001

subcontractor who NHTSA had, who I forget the name of that.
Q. Did Wecker work on that?
A. No, no.
Q. Okay.
A. I don't think I even knew Wecker at the time. I knew of him, but I didn't know him in the context of this case at all. Terry Klein was the guy at DOT who was involved in that.
Q. Okay.
A. There is another project that was at, not as major, but at the Bureau of Labor Statistics. Initial Multiple Imputation of Consumer Expenditure Survey. That involved Raghunathan and people at -- and me -- and people at the Bureau of Labor Statistics. Kathy Dippo, D I P P O; Geoffrey, with a G, Paulin, P A U L I N; Ralph Bradley, also at BLS, and some other people whose names I am afraid I don't remember right now.
Q. Okay. That is fine.
A. There is a project ongoing that

52299 4624

Produced by RJR in HUMPHREY/FTC
http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Raghunathan is involved in that I am not involved in, but it is through ISR, Institute for Social Research, at the University of Michigan, to multiply impute NHIS, N H I S.

Q. Which is what?

A. National Health Interview Survey.

Q. Okay.

A. The same one that is involved in these, that is being used in these cases.

Q. Okay.

A. I am not involved in that one at all.

Let's see if there are any others that come to mind. It is possible there are others I know about, and it is certainly possible that there are several others that I don't know about that are underway. That's close.

Q. Doctor, as I understand your testimony, the only multiple imputation project that you have not been involved with at all, even peripherally, is the Erasmus University one that you have named? Correct?

Produced by R.J.R.T.C. in MULTIPLE

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Support provided by the Industry Documents Library: www.industrydocuments.ucsf.edu/docs/hygl0001

MR. BIERSTEKER: I object to the form.

A. In fact, I was a slight advisor to him.

Q. Okay.

A. But I didn't -- I wasn't involved in any of the day-to-day stuff.

Q. Okay.

A. Nor am I involved at all in the NHIS project that is ongoing now, nor am I involved in the continuation of the Bureau of Labor Statistics project that Raghunathan is involved in. I forgot to mention that. I think it is ongoing.

Q. Okay. But you were involved originally? Correct?

A. I was involved in the first phase, but that is because I originated the idea, and now there are sort of other students who can do the work, and it is spreading out. There was a desire at the beginning, I think, to make sure the old man was involved in some sense.

Q. Okay. Now as I understand it then, was the output from phase one provided in a

Produced by RR Inc
Document in Industry

52299 4626

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.pdfpdf.industrydocuments.ucsf.edu/docs/hygl0001

rolling production to Wecker to do work on phase two, or did it have to be complete -- did phase one have to be completed before anything was sent to Wecker?

A. It basically had to be completed, but let me clarify that to make sure there is no confusion.

As I said earlier, when you are doing a multiple imputation, what you do is you randomly draw these things to complete one data set, and then you go back and do it again. So at the time, one data set was filled up by Raghunathan. Then that data set could be rolled over, in your words, to the Wecker guys while waiting for the second data set to be generated.

Q. How many data sets were generated by Raghunathan?

A. Well, we haven't completed it yet.

Q. So --

A. So the attempt would be to generate between five and ten.

Q. How many have been generated?

Produced by FTRFC
in the Public Interest

52299 4627

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

A. Well, we are not necessarily happy, fully happy with any of the ones that have been generated so far, but perhaps two have been generated that we -- we may be happy with.

Q. Two out of five to ten?

MR. BIERSTEKER: Well, objection to the form of the question.

A. Perhaps, but we don't know whether they satisfy the concerns that we have to make sure that they -- that there are no more bugs.

Q. All right. And I am going to ask you about the concerns you had about bugs or the output that you were getting. First I want you to complete your description of, if he -- have you completed your general description of phase two --

A. Yes.

Q. -- as to what work is ongoing?

A. Yes.

Q. Now how about phase three?

A. Okay.

Q. Is that started yet?

produced by R.J.R.T.C. in PHOTOPHRIA

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. Yes.

Q. And what is phase three?

A. Phase three is it takes the output of phase two and proceeds to allocate expense amounts that are associated with missing ICD 9 codes to the actual ICD 9 codes.

In the 18 ICD 9 codes that we are using and everybody is using, one of them is for missing, and I said there is no -- there is an event, and sometimes a dollar amount with an event, sometimes there is no dollar amount with the event, but NMES may categorize that event as having a missing ICD 9 code.

Q. Okay.

A. And so at the output of phase two, we use that 18th level of ICD 9 code as missing and have these events now that have expenses associated with them.

The third phase takes those events with missing ICD 9 codes but with now imputed expense amounts and allocates them to the 17 real ICD 9 codes. So at the end of phase three, we have one data

52299 4629

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

set that has no missing values for anybody on anything, and it has for the expense amounts it basically has for each of the ICD 9 codes it has real total expense amounts.

Q. There was a program required to be written for that?

A. Yes.

Q. Who wrote it?

A. I don't know who wrote it. I designed it. I described it to Gary Harvey and Wecker Associates, and we iterated on making sure what was going on was right, and then I think someone -- well, I don't think -- someone at Wecker Associates wrote it. I don't know who the person is. Gary Harvey might have. He might have assigned it to somebody.

Q. Now have you reviewed, yourself reviewed the output from phase one from Dr. Raghunathan?

A. I have reviewed pieces of output all along the way and in an attempt to detect problems.

produced in FTC HUMPHREY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. And have you detected problems?
A. Yes. In the past we detected problems. That is why we are still attempting to do it correctly.
Q. What were the problems you detected as to the output from phase one?
A. Oh, the kinds of -- the kinds of problems would be because of the coding problems -- I think in fact this was maybe in some report that I sent that was -- that is the one that held us up, I think, in last January, because the coding problems that differed between the Wecker NMES, as you referred to it, and the Raghunathan NMES, as you referred to it, that we found we were imputing people who had, for example -- this may not be exactly right, but -- people who were classified as nonsmokers as having a certain number of cigarettes smoked, and people who were nonsmokers as having a, oh, a quitting date, things like that, that don't make any sense, and that is because the coding is different. So the communication between the two data

produced by R.J. Reynolds Tobacco Co. in the LITIGATION LIBRARY

52299 4631

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the University of California, San Francisco Library
http://www.industrydocuments.ucsf.edu/docs/hygl0001

sets got confused and things being imputed that made no sense, that were just wrong.

Q. Did you detect any other problems?

A. There is the usual kind of modeling problems that occur, especially with small sample sizes and with a large number of variables. In order to do this correctly, it is critical to try to include as many predictor variables on the missing value as possible to maintain consistency of the data set.

But some of these expense categories are quite rare. Across the whole data set, there may have been only -- just approximately, for example -- 25 events of that particular type, that ICD 9 code in ambulatory or something like that. And of those 25, perhaps only 13 had observed expense amounts, which creates a small data set with observed values to try to predict something, especially if you are trying to predict it from 500 variables, approximately 500 variables.

Produced by RJ Reynolds in the UCSF Industry Documents Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

And so the coding of that had to be done very carefully to make sure that the kinds of values -- that the computer program wasn't going crazy in some sense trying to invert singular matrixes and trying to use overflows and underflows based on these computational problems with small data sets.

MR. BIERSTEKER: Off the record.

(Discussion off the record.)

BY MR. WITHEY:

Q. Is there such a thing as a diagnostic -- what is it -- diagnostician for all of these three phases?

A. Yes.

Q. Describe that process.

A. Well, that is basically the kind of process I am talking about, these sanity checks, are the values that are being imputed within the range of the data that are plausible? Are there errors that are occurring because you are trying to fit models that can't be fit? And so there is -- there are these diagnostic evaluations

Digitized by Tobacco Industry Documents

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

at each phase to make sure that the values that are being imputed look plausible and that the models that are being fit are being fit in realistic ways.

There are so many variables that each iteration -- that is why it takes so long. The computer is spinning, grinding away doing probably billions of multiplications every few minutes. And so it is -- you have to check to make sure there is not some switch that got switched the wrong way.

Q. Okay. So the diagnostic process is not part of any particular phase, but it is ongoing and has to be done on an ongoing basis as to all of the output?

A. Absolutely.

Q. Fair enough?

A. When you write computer programs or use computer programs, there are always bugs initially. You have to make sure that you have some faith in the answers at the end. So whenever you -- actually that is true when you are writing your own

52299 4634

produced by RJRTC
in Minnesota
http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

programs or even when using canned programs that are commercially available. It is very easy to misuse them and get results that make no sense.

Q. And I would assume that someone who is a diagnostician would have to really know the system you have helped create? Fair enough?

A. Absolutely. You have to know the system. You have to know the data. You have to be willing to iterate and go back and forth.

Q. For that reason, it is so time consuming?

A. Absolutely.

Q. All right.

A. Particularly here, because this phase one, if each run takes two or three days or two days, it is not like pushing a button, looking at the output, and saying, oh, now I know what I did wrong. Now you know what you did wrong, you have to fix it. It is another three days before you see whether you have fixed it. So the process is very, very cumbersome.

Q. And I assume that if we were to employ

Exhibit 52299 4635

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Digitized by the Internet Archive in 2014 http://www.industrydocuments.ucsf.edu/docs/hygl0001

some diagnostician or person who reviews it, it would be equally time consuming? Fair enough?

A. Hopefully at that point it will be rid of most of the bugs. But that is right. It would be -- it would be -- well, let me correct it. Actually not because -- let me go back.

If you want to run the software, that's correct. If you want to examine the let's say five multiply- imputed data sets that we give you, it would not -- you would have to be the diagnostician for those five, but it wouldn't require your rewriting the program and rerunning the program.

Q. Well, but it would still be time consuming? Fair enough?

A. Yes. It would be time consuming.

Q. And --

A. But much less so than writing and debugging a program. I mean an order of magnitude less.

Q. Let me ask you this then. Did you -- so

52299 4636

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

as I understand it then, did you also receive output as to phase two that you were the diagnostician for?

A. I was as people at Wecker. Yes.

Q. How about phase three? The same?

A. The same.

Q. Do you have an estimate of the total amount of time it has taken for you, Dr. Raghunathan, and Wecker & Associates on this project?

A. I have no idea how much Wecker Associates have, the time they have spent on it. I mean I can make a wild guess based on phone calls and --

Q. Go ahead.

MR. BIERSTEKER: Objection. Don't make a wild guess.

A. And Raghunathan, I would have to look at the records, although I haven't gotten any records from him for a couple of months, so I don't know what the last two months have been.

Q. How about yourself?

A. I would have to go back and look.

Legacy Tobacco Documents Library
UCSF Industry Documents Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. Give me your best estimate.
A. 100 hours. Maybe more.
Q. Would you assume that Dr. Raghunathan has more time in this than you have?
A. Yes.
Q. Would you assume Wecker & Associates has more time than you have?
A. Yes -- I think -- I would -- I would think so.
Q. All right.
A. But that might be wrong.
Q. Okay. That's your guess estimate?
A. Yes.
Q. That they have more time?
A. Yes. I think they have more time. Yes.
Q. Okay.
A. And my time might be off, off as well, I mean the number that I gave.
Q. And --
A. To help calibrate, I can tell you that the NCHS project went on for several years.
Q. Okay.
A. Computing is better now. We know more about it now.

UNOFFICIAL DOCUMENT

52299 4638
http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. Let me ask you this. On the other projects that you have been working on and have completed, including the ones that you mentioned, have you ever had anybody hired or asked to come in after the project was completed to review it and to essentially peer review it, if you will, or to look through it to see if there is any bugs, see if there are any obvious errors, try to determine if there was any methodological problems, review the output, kind of do some checks, sanity checks as you call them?

A. Yes, yes.

Q. Those kinds of things?

A. At the beginning of the question, I am not sure I am answering it right.

Q. Go ahead. Explain what you meant.

A. Because I -- in all of these major projects, there are -- we want there to be critics.

Q. Right.

A. And for the NCHS project, for example, the statisticians at NCHS who were involved in

Produced by IRRI in the UNIVERSITY of CALIFORNIA

52299 4639

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

helping us were critics who did a massive evaluation. That actually was the basis of their National Center for Health Statistics, their deciding to use multiple imputation in EDHANES when the old hot-deck methods that they regarded on the basis of our evaluations as unreliable relative to the multiple imputations that we created. There are several procedures, publications on that.

Q. And these, the massive evaluations done by NCHS, how long did that process take?

A. See, there is a -- I am hesitating, because there is doing the evaluations to see that it looks good, and then there is writing it up. Writing it up takes people's time when they decide to write.

Q. From start to finish including writing it up, how long did it take?

A. Well, if somebody is not working on it for a year, does that count? That's what I am saying. We did these evaluations. They decided to, this is the way to go, everything looks great. That was probably

Produced by R.R. Inc
Funded in part by

52299 4640
http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

about two years ago. And it still hasn't been fully written up. I am saying people got busy doing other things, although they decided this was the right way to go.

Q. How long did the evaluation process take?

MR. BIERSTEKER: I object to the form.

A. I am trying to remember. It was spread over several months for sure, because it wasn't a full-time activity of anybody's.

Q. Okay. Several months, meaning three or four, something like that?

A. Possibly. But again that is because people did it when they had time to do it, I think.

Q. Right.

A. And I don't -- I don't really know, because there was evaluation -- I was not involved in it. I tried to in fact kind of stay separate from it, because I wanted to be objective.

Q. Fair enough. Do you know if there was a similar evaluation performed on any of the other multiple imputation analysis

Produced in R.J.R.T.C. litigation

52299 4641

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Downloaded from www.industrydocuments.ucsf.edu/docs/hygl0001

projects that you have identified?

A. Yes. The DOT one, Department of Transportation one, the evaluations were done -- actually let me be clear about the evaluation.

Part of evaluation is when you are trying to debug programs and get it to work right. That is very time consuming. Once we are sort of happy with that, with that phase, then you go into the sort of final evaluation, can it answer really hard questions.

The first kind of evaluation is does it pass all of these sanity checks and get the bugs out. The next part is, okay, now let's pretend somebody really had this multiple-imputed data set and was using it to address the questions of interest, like a user, EDHANES, does it stand up to that.

Q. Okay.

A. That process is more straightforward. It doesn't involve this debugging. You think you have really done it.

52299 4642

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. The people that did that evaluation, as opposed to the debugging aspect of it, were those people that were familiar with the project?

A. Yes.

Q. All right. And thus they were aware of what was going on, at least generally, at the time the work was being done, and then they were brought in to do the evaluation, or are they part of the original team?

A. They were part of the original team.

Q. Okay.

A. Although they weren't part of the -- what might be called the hard computational part.

Q. Okay. And have you or do you ever know of an experience in which a multiple-imputation analysis has been done and then someone was brought in who was never involved with any aspect of the work until they were asked to evaluate it?

A. In some sense, yes. This project at the Bureau of Labor Statistics was interesting that way, in that at the time the Bureau

52299 4643

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

of Labor Statistics team was primarily economists. They were worried about this consumer expenditure survey and the amount of missing data in it on how people spend their money, their expenditures, and income. They were worried about income.

And in that project, Raghunathan and I were contacted, primarily Raghunathan originally, to multiply impute -- maybe it was me -- I don't remember who originally -- strike that --

Q. Well --

A. -- to investigate the possibility of multiply imputing the missing data in this consumer expenditure survey. The way it was divided up is we decided to create a real test case for that where we would create a data set at the Bureau of Labor Statistics that the staff there would then there create missing values in, give to us, meaning Raghunathan and me, to multiply impute, and give back to them to evaluate. So only they knew the right answer, because only they knew how the

Produced by the Tobacco Industry Documents Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Saved from http://www.industrydocuments.ucsf.edu/docs/hygl0001

missing values were created.

So they had nothing to do with the actual imputation itself. They were economists at the Bureau of Labor Statistics who were interested in finding out how well it worked. So they did the evaluation of it to find out how it worked relative to their methods and their suggestions. So in that sense, they were independent. Because we never -- we never knew the right answer.

Q. Well, I am talking about someone who would come in cold who has never even -- who was certainly never involved in the generation of or design of the computer programs that would have to be checked and analyzed, that was not involved with the generation of any of the data, wouldn't have to check that, and was not involved with the software programs that would be used, were not involved with the debugging of any of the problems that existed, but would have to come in fresh, anew, and try to evaluate this program. Have you ever had

produced by R.J.R.T. Industry Document Library

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Digitized by the Internet Archive in 2014 http://www.industrydocuments.ucsf.edu/docs/hygl0001

that experience?

MR. BIERSTEKER: Objection. It assumes facts not in evidence.

A. The Bureau of Labor Statistics project was really like that.

Q. All right.

A. Because the people evaluating it sat in on some meeting, because they were involved in the funding of the project, but they had nothing to do with any of the pieces of it.

Actually there is another one that is even cleaner. I don't know how the evaluations really went. It is the Federal Reserve example.

Q. All right.

A. The surveys.

Q. Consumer finance?

A. Yes. Consumer finance. That has been around I think close to eight years. It has been out there in the real world. There was a paper given on it last summer at the annual meeting of the Statistical Society. I think it has been evaluated by

produced by FTC in anticipation of litigation

52299 4646

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

a whole bunch of people involved in it.

Q. How long did that evaluation process take, if you know?

A. I don't know the people, so I don't know how long it took. But basically they are using a data set, and they are seeing how well it works. Apparently they think it works well. The evaluation doesn't have to be that involved.

Q. Okay. You could have fooled me.

A. Okay.

Q. And the Bureau of Labor Statistics, how long did that evaluation process that you have described take?

A. Oh, the final? Maybe a few weeks.

Q. Okay.

A. Oh, there is another example I can give you.

Q. Never mind.

A. An example that addresses exactly your question.

Q. We have to get this deposition done at five o'clock --

A. Okay.

Produced by an ITC approved reporter

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. -- or shortly there afterwards. Otherwise, Phillips and I won't be able to see our families tomorrow morning.

A. I was going to say it was the Census Bureau application.

Q. Is the work still in phase three or phase two, or which phase of the work is ongoing?

A. Well, we have all -- there are programs for all phases. We are in the process of still checking results, sanity checks, on the results of all phases.

Q. Was there any problems with the output that was generated from phase two that you had to deal with?

A. Phase two and phase three are much more straightforward, because they don't involve this iterative simulation, Markov Chain Monte Carlo stuff. Initially there were. At present I think they are fine.

Q. What were the initial problems?

A. Oh, things like putting in something upside down, putting in the inverse of something rather than -- dividing by

Produced by R.J.R.C. in the University of Minnesota Libraries

52299 4648

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the National Library of Medicine http://www.industrydocuments.ucsf.edu/docs/hygl0001

something rather than multiplying by something, the kind of coding problems that always occur when you are writing code.

Q. You have referred to in the past to writing code or coding. How is -- or coding problems. How was that term used in connection with doing a multiple imputation of NMES?

A. Okay. I actually am using coding in two different ways.

Q. Describe both.

A. The first I talked about coding problems that led to our problems before was coding responses on NMES. In other words, if you are a smoker, are you a current smoker, you are coded one, and never smoker, you are coded zero. Where is the variable put down, how is that coded.

Q. You have described that?

A. Right. The other use of coding, and I apologize for this ambiguity, is writing computer code, writing programs. These are programming problems. Computer code

Produced by LTC
in LAW

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

has programming problems.
Q. What is the quantity of the output that has been generated to date, if you can describe it in either electronic or hard copy terms?
A. The quantity?
Q. Yes.
A. Of the sort of final product or just all the amount all the way along?
Q. The final product.
A. The final product will be a -- it is a five copies of NMES, each copy with all the values filled in, no missing data at all.
Q. What is the most -- for any -- and you did this for each individual; correct?
A. Each individual have complete data.
Q. How many megabytes of computer space would that occupy, the five sets?
A. Five times as much as one -- as NMES occupies.
Q. And how much does NMES occupy?
A. Well, how many people is it? I don't know how many megabytes. I would have to

produced by RJ Inc
in PHILIP MORRIS LIBRARY

52299 4650

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.pdfpdf.industrydocuments.ucsf.edu/docs/hygl0001

multiply that out. It depends on how the individual variables are coded and stored. There are different ways of doing that. And I don't have any idea whether it is packed or not packed.

Q. What does packed mean?

A. Well, if I have a number, right, are you a smoker or not, that is a zero-one. Do I have to take up 32 bits to store zero-one? No. I don't. It is one bit.

Q. For any one individual, how many imputations -- what would be the maximum number of imputations done?

A. Well, right now we are planning to have a total of five data sets. That would be five. We may end up doing ten.

Q. Well, but you have described how oftentimes there is missing data for a whole series, a number of variables --

A. Right.

Q. -- that are looked at. If there is 500 variables, might there be someone that you have implied data for 100 of those variables, for instance?

Produced by RJRTC
FOR IMPUTATION

52299 4651

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

A. They will be all -- a person will have -- let me start again.
On each of these data sets, like the five data sets we will produce, each person will have fully observed data.
Q. Yes. 500 or so variables?
A. Right. Exactly.
Q. But the question I have is did you keep track of how many of those, that data, is multiply imputed as opposed to that which was data provided under the underlying NMES data?
A. Absolutely.
Q. Okay.
A. There are indicators for which values are imputed and which values are real.
Q. Do you have a recollection as you sit here what the most number of values that were imputed for any given individual?
A. No. I don't.
Q. Approximately?
A. I don't.
Q. Could there be as much as 100 to 200 variables that were imputed?

52299 4652

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

A. The only way you could ever get that high is if all the expense data, so if somebody had many, many events, medical events that were missing and each cell we had -- that -- that question is not -- is not quite -- it is hard to answer.
Q. It is not precise?
A. It is a good question, so don't -- don't make mistake. Because all you care about is for a particular type of expense and an ICD 9 code, you need to know the number of events and the total expenses in order to do the modeling, and so the imputation basically in the earlier phases is taking place at the event level. But then it is totaled to get a total number of expenses for those events. So we may be imputing hundreds of events, but it won't translate into hundreds of missing values for that person because they get totaled for that type of event.
Q. I understand that. There is a -- okay.
So when are you going to have those done?

produced by RJRTC in HUMPHREY

52299 4653

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

A. Well, are you going to help us do the diagnoses? I am kidding.
Q. No.
A. Seriously, we are in the process now of we have two of these five that we think are pretty good, and they are undergoing diagnoses now at the same time that we are generating more in hopes that we have gotten rid of all the bugs along the way.
Q. So when are you going to have it done? Mr. Biersteker may have asked you this question previously.
A. Yes. He asked me.
Q. I won't interfere with the dirty --
MR. BIERSTEKER: More than once.
A. And not nearly as politely as you do.
Q. Thank you very much. I was trying to be polite. But he is paying you, and I am not, so.
(Laughter.)
A. I think there is the hope that by the end of next week we will have five that we regard as acceptable.
Q. Did you have a conversation with

Produced by FRTC
in the UNITED STATES

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Support provided by the UCSF Library
https://www.industrydocuments.ucsf.edu/docs/hygl0001

Mr. Biersteker that went something like this, or any lawyer for the tobacco industry, that said -- went something like this: When is my report due?

It is due on -- the last report is due on May 26th, supplemental report.

Question: I won't have the multiple imputation analysis done by then.

And then a response saying -- well, no. Just end the question.

Have you had any conversation that you informed Mr. Biersteker that the multiple imputation analysis would not be done by May 26th or the date when your last supplemental report was submitted?

A. I'm sure we must have.

Q. Okay. And when you began this project in December, you were aware, were you not, or were you made aware, of the deadlines for the submission of reports?

A. I was made aware that there were deadlines. Absolutely. I didn't -- I don't know whether I knew the exact dates

Privileged & Confidential
Subject to Protective Order
In Litigation

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

for those deadlines, and I am assuming I was made aware that it was a -- trying to meet those deadlines was important, and I also was aware that it was a major project to undertake.

Q. And you informed him of that, I assume?

A. Yes.

Q. That it was a major project?

A. But I also informed him that I thought I had it divided into these phases in such a way that they could, that these phases could be worked on in parallel in the sense of phase one working with the Raghunathan group and phases two and three working with the Wecker group.

Q. Do you understand that once this project is completed that both Dr. Wecker and potentially other defense experts that have relied upon NMES data will be expected to supplement their reports to the extent to which they want to utilize the work that you have done?

A. I basically understand that.

Q. Okay. Do you know that those include

Produced by RJRTC in Philip Morris Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

Dr. Brian McCall? Have you heard of his name?

A. Yes.

Q. Okay. Dr. Wecker for sure?

A. Yes.

Q. And do you know of any other experts that are essentially waiting the results of this project so that they can review the data and supplement their reports other than --

A. I certainly am eagerly awaiting.

Q. You are?

A. Yes. Do I know anyone else besides?

Q. Yes.

A. I do not believe I know anyone.

Q. Okay. I think that is the three. What is your understanding, Dr. Rubin, of the degree of precision required of plaintiffs' damage estimates in this legal proceeding?

MR. BIERSTEKER: I object to the form of the question.

A. I don't believe I have an answer to that. I don't think I have an understanding

Produced by R.J. Reynolds Tobacco Company in HOUSTON

52299 4657

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Powered by http://www.industrydocuments.ucsf.edu/docs/hygl0001

per se.

My objective is to try to produce valid point estimates and valid estimates of uncertainty, whatever they happen to be, from the data that are available.

Q. Would you agree that the appropriateness of an estimation methodology to a given situation depends upon the degree of precision required of an estimate?

MR. BIERSTEKER: I am sorry. Can I have that question read back? I am not sure I understand it, or maybe I just missed it.

Q. Does the appropriateness of an estimation methodology to a given situation depend upon the degree of precision required of the estimate?

A. I'm not sure I follow. I could try to paraphrase and --

Q. Well, let me try to clarify it --

A. Okay.

Q. -- and see what you don't understand about it.

In other words, you would agree

produced by R.J.R.T.C. in LITIGATION

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

that in the real world and including in public health or in law or in academia or in -- in peer-review publications there may be different degrees of precision required?

A. For addressing different kinds of questions?

Q. Correct.

A. Correct.

Q. All right. And some of them require 95 percent confidence interval, for instance, and others require more probable than not degree of precision? Fair enough?

A. I wouldn't phrase it that way, because that's not -- that's not quite the right way to phrase it, but I understand what you are driving at. I would agree with the general intention of your question, I think.

Q. And, therefore, would you agree that the appropriateness of an estimation method, a method of estimating a value, medical cost of smoking, for instance, might depend

Produced by FTC
in HUMPHREY

52299 4659

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

upon the degree of precision that is required of that estimate?

MR. BIERSTEKER: Objection to the form of the question.

A. I am still not clear what you mean. I think I understand that you are saying that a particular estimation procedure that may be satisfactory for one objective may not be satisfactory for another objective.

Q. Well put. I wished I asked that. I was going to offer that one earlier. I am sorry. I should have let you go. You would agree with that? Correct?

A. Yes.

Q. Okay. And, for instance, are you familiar with the fact that from time to time medical doctors after reviewing a patient's medical history and clinical history and reviewing the diagnostic techniques comes to the conclusion that, for instance, the patient's smoking has caused them to incur lung cancer -- get lung cancer?

Produced by RJ Reynolds Tobacco Company

52299 4660

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

MR. BIERSTEKER: I object to the form of the question. It assumes facts not in evidence.

A. Do I understand that a doctor may sometimes reach that conclusion?

Q. Yes.

A. I understand that sometimes doctors reach those kinds of conclusions.

Q. Okay. Would you also understand that when doctors reach that conclusion they may tell their patient things like, well, I'm not 100 percent sure, but I think more probably than not that your smoking contributed to your lung cancer?

A. I understand that doctors may say that. Yes.

Q. Okay. And you understand that often doctors --

MR. WITHEY: Strike that.

Q. Have you examined any of the data sets upon which studies on smoking and health other than the Cancer Prevention Study II have?

MR. WITHEY: Strike that. Let me

52299 4661

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

go more specifically to your report.

Q. Do you have a copy of your report with you?

A. I can dig it out somewhere, but I don't have it.

Q. I think I may have a copy.

A. Do you have it in front?

Q. Let me just ask you to --

MR. BIERSTEKER: Are we going to get some free work product again, Mike?

MR. WITHEY: I don't think so. Not today.

MR. BIERSTEKER: Darn.

Q. The report I am referring to is the --

A. The First Supplemental?

Q. Yes.

A. Okay. Page 19. Okay.

Q. I don't need to mark your CV. I think we have already talked about that in the last deposition. But -- nor am I going to ask you questions about the first report, because we have already deposed you on that, I think.

But handing you what will be

Produced in HIRAY

52299 4662

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

marked as Exhibit 1, is this the supplemental report of you submitted around February of 1999, Doctor?

A. It looks to be.

(Document headed Northwest Laborers Litigation Supplemental Report of Professor Donald B. Rubin marked Exhibit No. 1 for identification.)

MR. WITHEY: Mark this as 2.

(Second Supplemental Report of Donald B. Rubin - Northwest Laborers marked Exhibit No. 2 for identification.)

BY MR. WITHEY:

Q. What has been marked as 2 is your second supplemental report submitted in May of 1999?

A. It looks to be. Yes.

Q. And what additional work did you do in this case to generate Exhibit 1 other than what you had previously done before?

A. Okay. My -- my memory is what I did to generate the supplemental report is I

Produced by R.I.R.T.C. in HUGHEY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Library
http://www.industrydocuments.ucsf.edu/docs/hygl0001

actually clarified some of the parts that were written in the original report and added some various analyses. I think they are primarily propensity score analyses or maybe entirely propensity score analyses that were not there in the initial report.

Q. Right.

A. I think also that, if I remember correctly, the supplemental report also addressed Dr. Harris. I don't believe he was in the first report.

Q. All right.

A. So I think there was some minor adjustments to deal with that.

Q. All right. And then Exhibit 2, what additional work did you do between February and May to generate the second supplemental report, Exhibit 2?

A. Okay. The primary work that is reported are propensity score analyses comparing CPS II with national representative databases.

Q. On page 19 of Exhibit 1, you list three conditions?

Produced by IE, TC in UNLIPHRAY

52299 4664

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. Correct.
Q. And those three conditions relate to the propensity scoring? Correct?
A. Right. Propensity scoring analysis to see how far apart groups are, treatment control groups are. Yes.
Q. The question I have for you is have you examined the data sets upon which other studies of smoking and health besides CPS II -- let me go back.

Have you examined the data sets upon which other studies of smoking and health besides CPS II were based to determine if any of the three conditions listed on page 19 of your February report exist --

MR. BIERSTEKER: Objection.

Q. -- or were satisfied I should say?

MR. BIERSTEKER: I object to the form.

A. Well, if you turn to page 21, I am doing these analyses for NMES and NHIS, N M E S and N H I S, so those are not CPS II, and they all appear to have, all meaning I

Produced by R.J.R.T.C. in HOUTHRAY

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Images at www.industrydocuments.ucsf.edu/docs/hygl0001

have got subgroups, males, females, formers and currents. They all seem to have fairly substantial biases between smokers and never smokers that suggest that relying entirely on models with their linearity assumptions is dangerous.

MR. WITHEY: Actually I am going to move to strike as nonresponsive. No offense. I need to probably state the question more clearly.

Q. My question is NHIS does not look at the issue of smoking and health outcomes, does it? I mean it doesn't to try to determine relative risk; correct?

MR. BIERSTEKER: I am going to object. Objection to the form of the question.

A. I'm not sure what you --

MR. BIERSTEKER: Can I help? Do you want me to help a little bit?

MR. WITHEY: Yes.

MR. BIERSTEKER: A data set doesn't really look at something. It is the analyst that comes behind.

Produced by the Tobacco Industry Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Available at http://www.industrydocuments.ucsf.edu/docs/hygl0001

THE WITNESS: Yes.

MR. BIERSTEKER: But go ahead.

MR. WITHEY: I said the data set which studies the relationship of smoking and health. Maybe that is why.

MR. BIERSTEKER: That is why --

MR. WITHEY: That is why I tried to ask it again. I don't think I asked it the right way.

THE WITNESS: Yes. What is bothering me is what --

BY MR. WITHEY:

Q. You understand there are studies of smoking and various disease outcomes that are in the epidemiological literature; correct?

A. Yes.

Q. You understand they are subject to biostatistical analysis; correct?

A. Correct.

Q. You understand they are based upon data sets that are generated in those studies or that are referred to?

A. Correct.

52299 4667

Produced by the University of California San Francisco Library in 2007 http://legacy.library.ucsf.edu/tid/ooq07a00/pdf

Digitized by the Internet Archive http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. Now as to those -- and there are probably hundreds or plenty of such studies on smoking and human health? Fair enough?
A. Fair enough.
Q. All right. And as to -- well, how many of those studies have you applied the three conditions listed on page 19 of your February supplemental report, if any?
A. I don't think you mean applied. You mean assessed.
Q. Assessed.
A. I imagine none, because I don't think I have ever looked at those other data sets. I am hesitating slightly, because I -- I may have through -- through a student's work, but I don't know.
Q. How many studies, epidemiological studies of exposure to other potential risk factors believed to cause disease have you assessed the three conditions on page 19 of your February report, if any?
MR. BIERSTEKER: Objection to form.
A. Epidemiological risk factors? I guess a

Produced by FTC
in litigation

52299 4668
http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

few.

Q. Which? Which risk factors and which diseases?

A. Well, I am thinking primarily now of work I did on prenatal exposure to barbiturates and hormones as to exposure in utero, where they were exposed and unexposed, and those analyses go back 15, 20 years ago initially when we were doing them. So we did these kinds of analyses.

I don't remember from that time whether I examined exactly these three criteria as listed here, but certainly they were -- they were there, because the work it refers to was done before that. It was done in the '70s.

Q. Well, --

A. So.

Q. -- I am specifically referring to those three criteria in this question.

A. Well, certainly, yes. Those were looked at for sure.

Q. If your three criteria were used as a screen to determine if the results of

Produced by the University of California, San Francisco Library in 2018

52299 4669

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Library
http://www.industrydocuments.ucsf.edu/docs/hygl0001

regression analysis are accepted for publication, what fraction of the papers based on regression analysis in the last say 10 years would have been published, do you believe?

A. I don't know. I really don't know.

Q. Do you think it is more than half?

A. I don't know, because it depends upon how -- how badly biased they are. I certainly know there are examples where these kinds of propensity scoring techniques would have prevented -- I don't know about epidemiology necessarily -- but would have prevented the publication of results that aren't particularly good.

Q. Have you ever analyzed a data set in a published epidemiological study and found it met all three of these criteria with respect to exposure at issue in that study?

MR. BIERSTEKER: Objection to the form of the question.

A. So have I ever seen an epidemiological study where I compared the background

Produced by in PHYSICAL

52299 4670

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

characteristics of the units on important characteristics and found it such that it was -- that the assessment was perfectly fine to go ahead and use regression models?

Q. Yes.

A. There was one that was sort of close that I was involved in regarding treatment of melanoma. It is not quite exposure we are looking at. It is comparing two kinds of treatments for stage 3 and stage 4 melanoma where the people did not differ that much as to background characteristics.

Q. Where was that published?

A. I'm not sure it was published as such. It was part of an FDA submission, part of the John Wayne Institute.

Q. My question was published.

A. I don't know if submissions to the FDA are considered publications. It may be.

Q. Is it publicly available in the medical literature?

A. I imagine it is, but I can't say for sure.

produced by R.J.R.T.C. in the LIBRARY

52299 4671

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. Can you think of any others other than that one?
A. Where I have seen where they are close, close enough to not be concerned?
Q. Well, that met all three of your criteria?
A. I don't.
Q. All right.
A. That is not the literature that I regularly read and then try to poke at to criticize.
Q. All right.
A. So no. I don't. I don't offhand.
Q. Now are you familiar with -- do you know Dr. Jim Robins, Harvard School of Public Health?
A. Jamie is what we call him.
Q. James. Sorry.
A. James or Jamie.
Q. And Steven Mark? Do you know those people?
A. Steven Mark, I know the name, but I am trying to -- I probably do know him, but I don't know him particularly well.
Q. Are you familiar with the article that was

Produced by the UCSF Library in HIV/AIDS HISTORY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Powered by the Legacy Tobacco Documents Library at http://www.industrydocuments.ucsf.edu/docs/hygl0001

published in the journal Biometrics, Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders?

A. I have not read it. No. But I -- I saw it referred to.

Q. By Dr. Harris?

A. Yes.

Q. Were you curious about what the article said?

A. Yes.

Q. I am going to satisfy your curiosity --

A. Good.

Q. -- since you haven't read it yet, I take it. I have not read it. I have not seen it.

MR. WITHEY: Mark this.

(Article entitled Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders marked Exhibit No. 3 for identification.)

BY MR. WITHEY:

Produced by R.J.R.
in HUMPHREY

52299 4673

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the UCSF Library www.industrydocuments.ucsf.edu/docs/hygl0001

Q. If you could read the abstract. I know I am not going to ask you to read the whole thing. If you can't answer the questions without reading the whole thing, let me know, because we may have to move on, but was a type of propensity.

MR. BIERSTEKER: Let him at least read the abstract before you ask the question.

THE WITNESS: Let me know where he is headed. Sure.

MR. WITHEY: That is exactly -- tell him to shut up.

MR. BIERSTEKER: I withdraw my objection.

MR. BIERSTEKER: Let the record reflect Mr. Withey had a smile on his face.

MR. WITHEY: I absolutely did.

THE WITNESS: All three of us did. Four.

BY MR. WITHEY:

Q. What I want to know is the following. One is the type of propensity scoring if any

52299 4674

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

used in the analysis, any disagreements with the authors about what they did, and what happened to the association between smoking and the FEV-1 that was looked at.

MR. WITHEY: While you are reading this, we can take a brief break.

MR. BIERSTEKER: Read as much as you like.

MR. WITHEY: Well, I don't --

MR. BIERSTEKER: I understand.

MR. WITHEY: I don't think he needs to read the whole thing.

MR. BIERSTEKER: I don't know that he can answer all of those questions just based on the abstract.

(Recess taken at 4:49 p.m.)

(Recess ended at 4:42 p.m.)

BY MR. WITHEY:

Q. Tell me when you are ready, Doctor.

A. We can try. I certainly understand the abstract, where it is going in generality.

Q. Let me just direct a question.

A. Sure.

Q. A type of propensity scoring method was

Produced by FTC in HUMPHAY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: http://www.industrydocuments.ucsf.edu/docs/hygl0001

used --
A. Yes.
Q. -- to adjust for confounding; correct?
A. It appears that way. Yes.
Q. And what were they used for?
A. It was used to adjust for multiple confounders in a situation where there are more than two treatment conditions, and that's what the generalization they say is beyond what Paul Rosenbaum and I did. That is one of the ways they say they generalize it.
Q. Do you have any disagreement with what the authors did with the propensity scoring?
A. In order to get into that, I would have to actually read the article far more carefully than I have in the last minute.
Q. Can you confirm from looking at table 1 and table 2 that the association between smoking and forced expiratory volume in one second, known as FEV-1, after these methods were used, did the association change?
A. I can't determine that. I am looking at

Produced by R.J. Reynolds in UNIPHIRE

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Support provided by the UCSF Library www.industrydocuments.ucsf.edu/docs/hygl0001

table 2 to try to see. Table 1 has just listed the variables. And I do remember that's what Harris said in his supplemental -- second supplemental report.

As I look at table 2, what it seems to be is -- and maybe this is incorrect -- but it seems to be where you do no adjustment at all, and then what you do, you do adjustment, use the propensity score methods, bringing in one or more variables into the propensity score estimation, but my memory of what Harris said and what you are saying is that these -- you are saying you initially do this by model-based adjustment, and then you do it by propensity, and the effect got bigger. And I don't see that at all. Maybe I am misreading it.

Q. So doesn't the negative relationship between smoking and FEV-1 increase in magnitude the from minus .0580 to minus .1133?

A. Yes. After I read the table, after my one

52299 4677

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

minute read of the table, which is not much time, is all they are doing is bringing more variables into the propensity scoring estimation. They are not comparing regression adjustment with propensity score adjustment. So I don't see why it is relevant. I mean as you bring in more -- you would believe the bottom result more, but that is just as you would if a model were correct you would believe the model more the more confounders you bring in.

So what they are showing, I think, is in a situation like this with potential confounders you should adjust for all the confounders you can and not just for something simple, like chronic cough or pack years of smoking or something like that, but you are better off adjusting for all the confounders.

Q. Let me see that back.

A. Okay. Maybe I misread it.

(Handing Exhibit No. 3 to Mr. Withey.)

Archived by the tobacco industry in LIBRARY

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. Or maybe I misread Harris' comments. That's possible, too.
Q. Table 1 lists 22 potential confounders to the effect of current smoking on forced expiratory volume in one second; correct?
A. That's what it says. Yes.
Q. All right. And these include such things as body mass, chronic cough, former pipe smoker, current asthma, et cetera? Correct?
A. That's what it says. Yes.
Q. And then on table 2, they -- I want you to see if you agree with me -- that there is an analysis of what is known as a constant term only and an analysis of a constant with the 22 covariates in table 1 that is listed. Did I read that correctly?
A. Correct.
Q. And I take it then that you don't -- without reading the article, you can't say what the significance of the application of these 22 covariates in table 1 in this table means as it relates to item number one, the constant term only; correct?

Produced by FTC in HOUPHREY

http://legacy.library.ucsf.edu/tid/oqf07c00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. I can make a good guess. They are referring to an equation which is for this the probability, the propensity score. It is basically saying that as you bring in more and more terms into the propensity score, as you are adjusting for more and more confounding variables by the propensity score, the answer gets -- changes, and it goes down to -- the answer you believe most is the bottom line, which is the lines that adjusts most of the covariates.

Q. Right.

A. It doesn't say anything about whether -- how that answer would compare to an answer based on a model-based adjustment. I see in this table no model-based adjustment at all, the usual sense of model-based adjustment.

Q. Okay.

A. Just as if you believe you like the bottom line you should bring, which I think I should in such a case, you should adjust for more and more confounding variables,

Produced by ... in the UCSF Industry Documents Library

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Digitized by the Internet Archive in 2014 http://www.industrydocuments.ucsf.edu/docs/hygl0001

and using propensity scale is a good way of doing it. That is my read based on 60 seconds.

Q. Dr. Harris in reviewing this article, which admittedly you haven't finished reading yet or started reading?

A. That's right. I haven't started.

Q. Methods based upon the propensity scoring methods actually strengthen the association between smoking and an FEV-1?

A. Strengthen relative to what?

MR. BIERSTEKER: Right.

Q. Well, between --

A. No adjustment at all.

Q. -- no adjustment at all?

A. Well, okay.

Q. Okay.

A. Not relative to model-based adjustment, which is what I thought the implication of his comment. Was he was talking about my criticism of relying on model-based adjustments alone? I was saying you should use both propensity and model-based adjustments to get reliable numbers.

Produced by R.J.R.T.C. in HOUDPHRA

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. I was asking about propensity..

MR. BIERSTEKER: Let him finish.

MR. WITHEY: I am sorry.

A. This article as far as I can tell does not address that whatsoever. So I don't see how that -- Harris' comment, now that I read the article, is relevant to the point that I thought he was making.

Q. I think the point made was sometimes when you do a propensity scoring method you may find that it is associated -- that the association between smoking and a given disease outcome may in fact be strengthened.

A. Strengthened relative to not adjusting for any confounding variables. It could happen. My interest is in getting the right answer.

Q. For all you know the application of that propensity scoring method to the data on relative risk in CPS II might also strengthen the association?

A. It could.

Q. Now are you familiar with either Kenneth

Provided by the Philip Morris Inc. Document Management

52299 4682

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.pdfpdf.industrydocuments.ucsf.edu/docs/hygl0001

Rothman or Sander Greenland?
A. Yes, I am.
Q. Those are pretty well world-renowned experts on epidemiology or professors of epidemiological, Dr. Rothman at the Boston University School of Public Health; correct?
A. Correct.
Q. Do you know them personally?
A. Yes.
Q. I assume you are familiar with their textbook Modern Epidemiology?
A. I know it exists. I have only glanced at parts of it.
Q. Let me read you something I just read to Dr. Mundt and see if you agree with both Drs. Rothman and Greenland as well as Dr. Mundt as follows. Let me read it to you. This relates to the issue of attributable risk formulas or attributable fractions that you have talked about in your paper --
A. Okay.
Q. -- or in your report, I mean.

52299 4683

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

"There is a tendency to think that the sum of the fractions of disease attributable to each of the causes of the disease should be 100 percent. It is, however, neither a difficulty nor an anomaly but simply a consequence of allowing for the fact that no event has a single agent as the cause. The fraction of disease that can be attributed to each of the causes of the disease and all the causal mechanisms has no upper limit. For cancer or any disease, the upper limit of the total of the fraction of diseases attributable to all the component causes of all the causal mechanisms that produce it is not 100 percent, but infinity. Only a fraction of disease attributable to a single component cause cannot exceed 100 percent."

Let me show you the paragraph, page 13, and ask you if you agree with that.

A. The first full paragraph?

Q. That's correct.

Produced by R.J.R.N.C. in the Tobacco Industry Documents Library

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Digitized by the Internet Archive in 2014 http://www.industrydocuments.ucsf.edu/docs/hygl0001

(Handing documents to the witness.)

(Pause.)

A. I have now read it.

Q. Do you agree with it?

A. The "but infinity" is kind of dramatic.

Q. True enough.

A. You have to have -- how many causes do you have to have, each one itself is bounded by 100 percent, how many would you have to have to reach infinity? Dramatic.

Q. How about can certainly can exceed 100 percent? Fair enough?

A. Yes.

Q. You agree with that?

A. Yes.

Q. All right. Have you read the report of Dr. Mundt in this case?

A. No. I have not.

THE WITNESS: Off the record?

MR. WITHEY: Yes.

(Discussion off the record.)

MR. WITHEY: Back on the record.

THE WITNESS: May I take a

52299 4685

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

two-minute break while you are looking that up?

MR. WITHEY: If you need to.

THE WITNESS: I need two minutes.

MR. WITHEY: Then we are going to go after five.

THE WITNESS: You will be done at two minutes after five.

MR. WITHEY: Fine with me.

(Recess taken at 4:54 p.m.)

(Recess ended at 4:56 p.m.)

MR. WITHEY: Back on the record.

BY MR. WITHEY:

Q. I want you to read page 15 of Dr. Mundt's report starting "Interpretation of attributable fraction measures." Just read the first two paragraphs, if you could.

(Handing document to the witness.)

(Pause.)

(Witness examining document.)

A. Okay.

Q. Okay. Read aloud the last sentence from

Produced by RJ Reynolds in LITIGATION

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

the second paragraph, please.

A. I was puzzling about that as well.

"Thus, it is erroneous to interpret a single AFM as the proportion of disease which would be eliminated upon removal of a risk factor, which would require that the sum of all such fractions not exceed 100 percent."

One of the things that makes it hard to, for me to interpret that last sentence, is that there is some words and uses of words in the earlier part that I was reading that I don't -- I'm not -- I would have to read the beginning part to know what he really meant by that. So it is taken out of context.

Q. Does that sentence appear to follow from the preceding part of the paragraph, the sentence you just read out loud?

A. I'm not sure. That's -- that's --

Q. Fine.

A. I'm not quite sure.

Q. Was the --

A. It could, if I had the context better,

produced by R.J.R.T.C. in the interest of

52299 4687

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Digitized by the Internet Archive http://www.industrydocuments.ucsf.edu/docs/hygl0001

but.

Q. Does that paragraph accurately reflect what Rothman and Greenland are saying in the portion of the book that I just showed to you?

A. The fact that he quotes that sentence from Rothman and Greenland gives me the feeling that he thinks he is writing consistently with it, so that's why I think that I better get the context --

Q. All right.

A. -- more complete to understand if it is consistent.

Q. At least it is not clear to you that that last sentence is in fact -- follows or accurately reflects what Rothman and Greenland were saying?

A. Well, those -- those two sentences, there is an intermediate sentence that could be missing, or that intermediate sentence could be filled in by reading the rest of the document.

Q. Well, let me ask you this. Is it proper to interpret an attributable risk value as

Produced by R.J. Reynolds in litigation

52299 4688

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

the proportion of disease which could be prevented upon the removal of that risk factor?

A. It can be.

Q. Okay.

A. That doesn't mean it is, but it can be.

Q. And that requires a judgment of epidemiologists or others to make? Correct?

A. More -- that and other things. That's what this paper that was included at the end of one of my supplemental reports was trying to do, to formulate that precisely.

Q. Now which paper was that?

A. That was the one in the Proceedings of the American Statistical Association.

Q. Could you name the paper from your -- what you have there?

A. I don't know.

(Pause.)

(Witness examining exhibits.)

A. Was it ever included as a supplement in Northwest? I don't remember now.

Q. I don't know either.

Produced in RTF

Produced in Industry

52299 4689

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

A. It was included in one of the other cases that you and I met on, so maybe it was Ohio.

Q. Look and see if it is in there.

A. It is not included in this report.

Q. That is Exhibit 1. How about in Exhibit 2? It might have been included?

A. No.

MR. BIERSTEKER: Wait, wait, wait.

THE WITNESS: What is it?

A. An overview. On page 15.

MR. BIERSTEKER: It is cited. It is not clear to me whether it was attached or not, Michael.

MR. WITHEY: Okay.

THE WITNESS: Cited at the bottom of page 15 of the first supplemental report.

MR. WITHEY: Let me take a look.

THE WITNESS: Sure. I will give you the whole thing, so it doesn't get separated.

BY MR. WITHEY:

produced by R.J.R.T.C. in HUMPHREY

52299 4690

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

Q. Page 15?
A. Yes. The bottom of it.
MR. WITHEY: Have we been provided with that, Peter?
MR. BIERSTEKER: Yes. If you want another copy, I can send it to you.
MR. WITHEY: No. That is okay. Okay.
BY MR. WITHEY:
Q. I think it was in your second supplemental report you discussed the fact -- maybe it was in your first -- that CPS II population was more homogenous than the national population?
A. Correct.
Q. And by that I assume you mean the difference between smokers and nonsmokers were not as different in the CPS II as in the general population?
A. No. I didn't -- I didn't mean that. Let me take a look.
What I was saying is that if you looked at just the general spread of the distribution of variables in CPS II, they

Provided by the Tobacco Industry Publications Archive

52299 4691

http://legacy.library.ucsf.edu/tid/oqf07a00/pdf
Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

were less spread out. So, for example, the age distribution was more compressed.

The other -- the other -- what other variables? Age, race, marital status, education, body mass index, high blood pressure and diabetes, were -- were more compressed in CPS II than in these national representative samples.

Q. Well, would you believe that that homogenous quality would apply to the differences, if any, or the -- yes, between smokers and nonsmokers in that population?

A. Not necessarily.

Q. Okay. So you don't have an opinion more probably than not that the differences between smokers and nonsmokers in the CPS II would be less than the differences between smokers and nonsmokers in the general population?

MR. BIERSTEKER: I will object to the form of the question.

A. Because CPS II is more compressed with respect to variability, then it is likely

Produced by R.J. Reynolds in litigation

52299 4692

http://legacy.library.ucsf.edu/tid/oqt07a00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

that it -- it -- it would be more compressed with respect to variability between smokers and nonsmokers.

Q. That is kind of what I thought. So now if that's true -- okay. No. Actually that answers the question. Let me see if there is anything else.

I assume by more compressed, you are referring to more homogenous?

A. Correct.

Just to clarify, if you can't -- if you have a data set where all the ages are between 35 and 55, it is hard to get smokers and nonsmokers more different than 35 and 55.

Q. That is a good example.

A. Whereas if the data set varies between 20 and 80, then it is possible smokers and nonsmokers differ by 20 to 80. Not very likely that much, but you get the picture.

Q. And does the degree of compression of those variables, or the homogeneity, tend to support the internal validity of the

Produced by R.J.R.T.C. in HOUSTON

52299 4693

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

relative risk derived, everything else being equal?

MR. BIERSTEKER: I object to the form of the question.

A. If -- I am supposed to suppose now there is more overlap in the smokers and nonsmokers in CPS II than in NMES, for example? Is that what you are saying?

Q. No. I am talking about the internal validity. More than the general population. Let's put it that way. That it is more -- that the differences between smokers and nonsmokers is less in the CPS II study than in the general population?

A. Okay.

Q. If you want to use NMES, I suppose, but.

A. I am supposed to assume that that is true?

Q. Yes.

A. Right. I assume that is true. Now go on. I am sorry.

Q. The question is, therefore, would that fact, everything else being considered,

produced by RJRTC in HUMPHREY

52299 4694

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Downloaded from www.industrydocuments.ucsf.edu/docs/hygl0001

render the relative risk derived from CPS II more reliable?

MR. BIERSTEKER: Objection to the form.

A. No. What it would mean is that if you derive the relative risk from CPS II using models with lots of confounders brought in that those models would be less sensitive to the modeling assumptions than they would be in a data set where there are bigger differences between smokers and nonsmokers.

Q. Would it mean that there would be -- would it mean -- would the lack -- excuse me -- would the relative homogeneity of the CPS II population mean, again everything else being equal, that the relative risks are less likely to be confounded?

A. Not really. It is just that the -- that the -- the ability to eliminate the confounding, it is easier to eliminate the confounding, and let me clarify. Everything else being equal -- you put everything else being equal in. So if I

Produced by Tobacco Industry Document Library

52299 4695

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

think about that in a broad enough way, then I think the answer to your question is probably yes, because they are less -- there are less biases, and there are more -- but it is only because we are looking at it at the wrong population subset. It is a little piece of the population where things are simple -- simpler to do than with the right population.

Q. Okay. Great.

MR. WITHEY: Thank you, Doctor. I have nothing further.

THE WITNESS: Sure.

MR. WITHEY: Obviously, I think we should state for the record -- it doesn't need to be stated -- that we in no way have waived our right to object to and move to exclude any testimony based on any multiple imputation analysis that hasn't been completed yet. I don't know if I even had to say that, Peter. I am not trying to beat you up on a Friday night.

MR. BIERSTEKER: I don't think you needed to say it.

Produced by RRPTC in DIMPFLY

52299 4696

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
http://www.industrydocuments.ucsf.edu/docs/hygl0001

MR. WITHEY: I understand.

MR. BIERSTEKER: I understand that that is your position.

MR. WITHEY: Thank you.

(Whereupon, at 5:06 p.m., the deposition was adjourned.)

produced for FTC in HUFFY

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Document hosted at www.industrydocuments.ucsf.edu/docs/hygl0001

DEPONENT'S ERRATA SHEET

AND SIGNATURE INSTRUCTIONS

The original of the Errata Sheet has been delivered to Peter J. Biersteker, Esq.

When the Errata Sheet has been completed by the deponent and signed, a copy thereof should be delivered to each party of record and the ORIGINAL delivered to Byers & Anderson Court Reporters, to whom the original deposition transcript was delivered.

INSTRUCTIONS TO DEPONENT

After reading this volume of your deposition, indicate any corrections or changes to your testimony and the reasons therefor on the Errata Sheet supplied to you and sign it. DO NOT make marks or notations on the transcript volume itself.

REPLACE THIS PAGE OF THE TRANSCRIPT WITH THE COMPLETED AND SIGNED ERRATA SHEET WHEN RECEIVED.

produced by RJRTC in HUMPHREY

52299 4698

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

CERTIFICATE

Commonwealth of Massachusetts
Plymouth, ss.

I, Judith McGovern Williams, a Registered Professional Reporter and Notary Public in and for the Commonwealth of Massachusetts, do hereby certify:

That DONALD B. RUBIN, Ph.D., the witness whose deposition is hereinbefore set forth, was duly sworn by me and that such deposition is a true record of the testimony given by the said witness.

IN WITNESS WHEREOF, I have hereunto set my hand this 8th day of July, 1999.

Judith McGovern Williams

Judith McGovern Williams
Registered Professional Reporter
Certified Realtime Reporter
Certified Shorthand Reporter No. 130993

My Commission expires:
March 22, 2002.

produced by RJRTC in HUMPHREY

52299 4699

http://legacy.library.ucsf.edu/tid/oqo07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

ATTACH TO DEPOSITION OF: DONALD B. RUBIN, Ph.D.

CASE: NORTHWEST LABORERS/PHILIP MORRIS

ERRATA SHEET

INSTRUCTIONS: After reading the transcript of your deposition, note any change or correction to your testimony and the reason therefor on this sheet. DO NOT make any marks or notations on the transcript volume itself. Sign and date this errata sheet (before a Notary Public, if required). Refer to Page 113 of the transcript for errata sheet distribution instructions.

PAGE LINE

____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____
____	____	CHANGE: _____
		REASON: _____

I have read the foregoing transcript of my testimony and except for any corrections or changes noted above, I hereby subscribe to the transcript as an accurate record of the statements made by me.

_____

52299 4700

Produced in HUMPHRIES
http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Produced by the UCSF Industry Documents Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

Rubin
EXHIBIT NO. 1
7-2-99
J. WILLIAMS

NORTHWEST LABORERS LITIGATION
SUPPLEMENTAL REPORT OF PROFESSOR DONALD B. RUBIN

I am a professor of statistics at Harvard University. I served as Chairman of Harvard's Statistics Department for nine years, from 1985 to 1994. A copy of my most recently-prepared curriculum vitae is attached to this report as Exhibit 1.

I have been asked to analyze the estimates prepared by plaintiffs' experts, Drs. Harris and Dement and Mr. Roberts, of the excess health-care spending by the plaintiff union trust funds in the Northwest laborers litigation attributable to defendants' alleged misconduct.

In this Report, I incorporate all of my opinions from my original report dated November 6, 1998 and I report my additional opinions based upon my review of the new reports filed by the plaintiffs' experts, Drs. Harris, Dement and Roberts, in late December, 1998. Thus, this Report contains all of my principal opinions in this case.

In this Report, I express five broad opinions, each of which is explained more fully below:

1. Reliable and statistically valid estimates of the health-care expenditures, if any, incurred by the plaintiff trusts as a result of defendants' alleged wrongful conduct can be calculated.

2. The analyses of the plaintiffs' experts, Dr. Dement and Mr. Roberts, do not even claim to estimate the excess medical expenditures of the trusts due to the effect of the defendants' alleged misconduct at all, including, particularly, the excess expenditures of the trusts due to the effect of that alleged misconduct on the trusts' behavior.

3. It appears that the analyses of Drs. Dement and Roberts attempt to estimate the excess medical expenditures of the trusts due to the existence of smoking. But those estimates, confidence intervals, and any statistical significance claimed for those estimates provided by the plaintiffs' experts are unreliable and statistically invalid.

4. Although Dr. Harris' December 30, 1998 report does attempt to estimate the effect of defendants' alleged misconduct generally on the medical expenditures of the trusts, his analyses essentially have none of the characteristics that they must have to generate statistically valid estimates of

52299 4701

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

the trusts' expenditures, if any, that were incurred because of the defendants' alleged misconduct generally. Those estimates, confidence intervals, and any statistical significance claimed for those damage estimates provided by Dr. Harris consequently are unreliable and statistically invalid.

5. Dr. Harris does not even attempt to estimate the excess expenditures of the trusts due to the effect of the alleged misconduct on the trusts' behavior.

I. HOW TO ADDRESS THE QUESTION

A. Introduction

To address the question of what excess sums, if any, the trusts expended as a result of defendants' alleged misconduct, one must compare the health-care costs that the trusts actually incurred to what those costs would have been in counterfactual worlds without the alleged misconduct. To start, one has to specify precisely what the alleged misconduct was and when it occurred, as well as what would have occurred in the absence of that misconduct and when it would have incurred. Because there are no direct data on the effect of defendants' alleged misconduct on the trusts' excess costs, two models must be used: a behavioral model and a medical expenditure model.

The behavioral model would have one or two components, depending upon whether the trust funds can recover excess expenditures due to (1) the general effect of that alleged misconduct, including effects on the trusts' behavior and on the behavior of the individual participants in the trust, or (2) the effect of the defendants' alleged misconduct on the trusts' behavior.

The one-component behavioral model would estimate the effect of the alleged misconduct on the subsequent smoking behavior of individuals who were or would have been participants in the plaintiff trusts, regardless of whether that change in smoking behavior was due to changes in the trusts' behavior.

52299 4702

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

The more complicated behavioral model would have two components. The first would address "individual-initiated" changes in smoking behavior and the second would address "trust-initiated" changes in smoking behavior. The first component would estimate the effect of the alleged misconduct on the subsequent smoking behavior of individuals who were or would have been trust participants, assuming no trust- initiated behavior was inhibited by the alleged misconduct. I call this the "individual-initiated" component of the behavioral model because its focus is on the savings, if any, to the trusts, due to possible changes in individuals' smoking behavior caused exclusively by the effect on the *individual* of the lack of the alleged misconduct in a counterfactual world.

The second component of the behavioral model -- the "trust-initiated" component -- would estimate the effect of the alleged misconduct on trust behavior in further reducing the smoking of its individual members through, for example, preventative programs. I refer to this as the "trust-initiated" component of the behavioral model because it focuses on the savings, if any, to the trusts, due to possible changes in individuals' smoking behavior caused exclusively by the effect of some behavioral action taken by the trust as a result of the lack of the defendants' misconduct in a counterfactual world.

The second model -- the medical expenditure model -- would estimate the effect on the trusts' expenditures of those changes in smoking behavior estimated from the first model. The cost savings that a trust could have effected in a counterfactual world without the alleged misconduct are the reduced costs, if any, caused exclusively by trust-initiated changes in the smoking behavior of individual trust members.

There are several other essential features of the proper data collection and modeling approach to address the questions posed by this lawsuit. For example:

Produced by R.J. Reynolds in HUMIRRA

52299 4703

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Stored on www.industrydocuments.ucsf.edu/docs/hygl0001

1. Both models must control for important background and other confounding variables so that they compare behaviors and costs for like individuals, that is, for matching individuals in the actual world (with defendants' alleged misconduct) and in the counterfactual worlds (free of that alleged misconduct).

2. Both models must take into account the passage of time, beginning when the alleged misconduct occurred and continuing through each subsequent, relevant year.

3. Both models must focus on the population of interest. Here, that is those individuals who were or would have been recipients of health-care coverage from one of the trusts in this litigation.

4. Both models must consider distinct types of smoking behavior that can lead to different health-care expenditure outcomes.

5. For those aspects of either model that rely on assumptions rather than data, the assumptions must be explicated and justified. It is critical, moreover, that each such assumption be capable of being individually assessed and altered.

6. The possibility that smokers' other health-related behavior may be affected by the alleged misconduct must also be considered.

7. The statistical techniques employed in both models and in the studies upon which those models rely must be reliable and statistically valid. For instance, missing data must be addressed in an appropriate manner; adjustments for background and other confounding variables must be made after taking into account differences in the distribution of those variables in the groups being compared; and sound statistical methods must be employed to reflect the reliability and uncertainty in the resulting estimates.

8. The health-care expenditure model must consider distinct types of health-care expenditures that can be differentially influenced by smoking behaviors.

9. Health-care costs that would have been incurred in counterfactual worlds without the defendants' alleged misconduct must be accounted for when estimating damages. For instance, any increased health-care costs due to smoking that occurred before any alleged misconduct cannot be attributed to that alleged misconduct because those costs also would be incurred in counterfactual worlds without the alleged misconduct.

10. To the extent that the law limits the trusts' recovery to excess health-care costs incurred because of the effect of the defendants' alleged misconduct on

Produced by RJR
Confidential
52299 4704
4 of 37
http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Stanford Research Institute www.industrydocuments.ucsf.edu/docs/hygl0001

the trusts' behavior, then savings in health-care costs that would have occurred in a counterfactual world without the alleged misconduct and without any modified behavior on the part of the trusts (*i.e.*, individual-initiated changes in smoking behavior) must also be excluded. For instance, cost savings from individuals whose smoking would have decreased without the defendants' alleged misconduct, regardless of whether or not the trusts modified their behavior, must be excluded.

11. Finally, expenses incurred in instituting preventative programs would have to be deducted from any potential savings due to reduced smoking prevalence arising from such programs.

B. Behavioral Model

1. The one-component behavioral model

Consider first the model of the effect of defendants' alleged misconduct on the smoking behavior in the pertinent population. This model was described in Section I.B.1 of my original report. For completeness, it is also described here.

Suppose it were alleged that the defendants failed to disseminate information about the health effects of cigarette smoking in 1965. I would first want to estimate the effect of that alleged failure on the smoking behavior of the individuals who were or would have been the recipients of the trust-funded health care. For each distinct type of smoking behavior that is considered relevant to health-care costs, I would estimate its prevalence in the actual world and in the counterfactual world.

This estimation would be based on actual world data and on expert opinion concerning the effect on smoking cessation and smoking initiation of the availability of additional information about smoking's health risks. Such a model of smoking behavior would clearly have to consider background characteristics of the individuals who were or would have been recipients of the trust-funded health care; control for those background characteristics and other confounding factors that alter the effect of the alleged misconduct on smoking behavior, and consider how changes in

Produced by IRJRTC in LITIGATION

52299 4705

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Batches hosted at www.industrydocuments.ucsf.edu/docs/hygl0001

smoking behavior in the pertinent population would vary from 1965 through the end of the damage period.

Even when the objective of an analysis is solely to estimate aggregate quantities (*i.e.*, not within subgroups such as defined by year of birth and race), one must generally control for background characteristics (such as year of birth and race) and other confounding factors (such as nonsmoking health-related behaviors) in order for the resulting estimates and inferences to be valid and reliable.

The task of creating a behavioral model is not as daunting as it might appear: simple modifications of actual world prevalences of different types of smoking behaviors within specific subgroups of people lead to estimates under explicit assumptions.

2. The two-component behavioral model

Alternatively, the two-component behavioral model would permit the plaintiff trusts to estimate the excess health-care costs, if any, that they incurred as a result of the effect of defendants' alleged misconduct on the trusts' behavior. The second component of this model was described in Section I.B.2 of my original report, but is described in more detail here after describing the first component.

a. Individual-initiated changes in smoking behavior

Consider first the "individual-initiated" component of the model of the effect of defendants' alleged misconduct on smoking behavior in the pertinent population. Suppose, again, that it were alleged that the defendants failed to disseminate information about the health effects of cigarette smoking in 1965. I would first want to estimate the effect of that alleged failure on the smoking behavior of the individuals who were or would have been the recipients of trust-funded health care.

Produced by R.J. Reynolds in HOUSTON LIBRARY

52299 4706

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Additional text at www.industrydocuments.ucsf.edu/docs/hygl0001

assuming no changes in the trusts' behavior in the counterfactual world. For each distinct type of smoking behavior that is considered relevant to health-care costs, I would "estimate" its prevalence in the actual world and the counterfactual world.

This estimation would be similar to that described above for the one-component model. It would be based on data and expert opinion concerning the effect on smoking cessation and smoking initiation of the availability of additional information about smoking's health risks; it would consider background characteristics of the participants in the trusts; it would adjust for those background characteristics and other confounding factors that alter the effect of the alleged misconduct on smoking behavior; and it would consider how changes in smoking behavior in the pertinent population would vary from 1965 through the end of the relevant period. This is necessary to obtain estimates and inferences that are valid and reliable, even when the objective of an analysis is solely to estimate aggregate quantities (*i.e.*, not within subgroups such as defined by year of birth and race).

b. Trust-initiated changes in smoking behavior

Now consider the second component of the behavioral model, the "trust-initiated" component, which concerns the behavior of the trusts due to alleged misconduct. The model must supplement the previous analysis with a posited potential reduction in the prevalence of smoking, also disaggregated by background characteristics and other confounding factors describing the trusts' members, that the trusts would have caused in the absence of the alleged misconduct. Suppose again that it were alleged that the defendants failed to disseminate information about the health effects of cigarette smoking in 1965. I would want to estimate the effect of that alleged misconduct on the behavior of each trust, and then estimate the effect of any modified behavior by the trusts due to lack of alleged misconduct on the smoking prevalence of the individuals who were, or would have been,

52299 4707

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf
Document is available at http://www.industrydocuments.ucsf.edu/docs/hygl0001

the recipients of trust-funded health care. For each distinct type of smoking behavior that is considered relevant to health-care costs, I would again "estimate" its prevalence in the counterfactual world for members of each trust. The analysis would, in general, require a separate longitudinal model for each trust because any potential trust-initiated changes in smoking behavior due to lack of alleged misconduct could have varied across trusts and in time.

C. Health-Care Expenditures Model

As described in Section I.C of my original September 11, 1998 report, once I had modeled the effect of the alleged misconduct on smoking behavior over time, I would then model the effect of that changed smoking behavior on the health-care expenditures of the trusts' members.

This model of the effect of the changed smoking behavior on health-care expenditures of the trusts' members would also have to consider characteristics of the individuals who were or would have been recipients of trust-funded health care. These characteristics would include year of birth, sex, race, income level, education, baseline mental and physical health, and other confounding factors, such as health-related behaviors, that may be important predictors of how smoking behavior affects the medical costs of the trusts' members. As with the behavioral model, ideally each such characteristic would be measured on each individual in the data set used to estimate the health-care expenditures model before the moment in time when the alleged misconduct had an effect on that characteristic for that individual.

The health-care expenditure model also would have to take into account the passage of time because costs accumulate in time and smoking behavior can change in time.

Such a model may have some general similarities to the plaintiffs' experts' medical expenditure models presented in litigation brought by the various states against these same

Produced in LITIGATION

52299 4708

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Batches loaded from www.industrydocuments.ucsf.edu/docs/hygl0001

defendants, but without their errors, with implicit assumptions made explicit, and with proper consideration of the passage of time.

D. Illustrative Examples

The essential characteristics of the models necessary to address the trusts' expenditures incurred due to the defendants' alleged wrongful conduct can be illustrated by examples. These examples are meant to illustrate specific issues; they are not meant to suggest that the models need to operate at this level of detail. The examples were presented in Sections I.D.1 and I.D.2 of my original report but are presented here with some clarifications. Also, the discussion in Section I.D.3 of my original report is expanded here.

1. Example One

Four of the essential features of the modeling required here are illustrated by a basic example. First, the example demonstrates that, in order to estimate separately the effect of defendants' alleged misconduct on the trusts' behavior, one must consider the actual world with the alleged misconduct and at least two counterfactual worlds without the alleged misconduct, one without changed behavior on the part of the trust and another with possibly changed behavior on the part of the trust. In contrast, if one only wants to estimate the effect of the defendants' alleged misconduct in the aggregate, then one must consider the actual world with the alleged misconduct and only one counterfactual world. Second, it shows the need to consider background and other confounding characteristics. Third, it illustrates the need to consider the passage of time. Finally, the example illustrates that there is modeling uncertainty about what would have happened in a world without the alleged misconduct, and so assumptions and bases for assertions about counterfactual worlds without the alleged misconduct must be explicated.

Produced by R.J. Reynolds in HIGHLY CONFIDENTIAL

52299 4709

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Support provided by the UCSF Industry Documents Library. http://www.industrydocuments.ucsf.edu/docs/hygl0001

In this example, we consider five worlds, one the actual world as it exists and four counterfactual worlds without the defendants' alleged misconduct starting in 1965. Two of the counterfactual worlds have no behavioral modification on the part of the trusts, and in the other two, the trusts initiated smoking cessation programs. We consider the same individual from one of the trusts in all of the worlds, which obviates the need to control statistically for background characteristics and other confounding factors. We then see how this individual's health-care costs differ in time in these five worlds, which leads to the calculation of this individual's trust's "excess" health-care costs due to the alleged misconduct.

		Counterfactual Worlds: Defendants' Disseminate Information on Health Risks			
	Actual World: Defendants Withhold Information	Trust's Behavior Unchanged		Trust Initiates Preventative Program in 1970	
		Counterfactual World 1	Counterfactual World 2	Counterfactual World 3	Counterfactual World 4
1935	Born	Born	Born	Born	Born
1955	Starts smoking	Starts smoking	Starts smoking	Starts smoking	Starts smoking
1965	Defendants withhold information on health risks - continues smoking	Defendants disseminate information on health risks - quits smoking	Defendants disseminate information on health risks - continues smoking	Defendants disseminate information on health risks - quits smoking	Defendants disseminate information on health risks - continues smoking
1970	Still smoking	Quit for 5 years	Still smoking	Trust initiates smoking prevention program - quit for 5 years	Trust initiates smoking prevention program - quits smoking
1985	Quits smoking	Quit for 20 years	Quits smoking	Quit for 20 years	Quit for 15 years
1990	Quit for 5 years	Quit for 25 years	Quit for 5 years	Quit for 25 years	Quit for 20 years

First, consider counterfactual world 1 where there are individual-initiated changes in smoking behavior effected by the lack of the alleged misconduct. There, the increase in the individual's

52299 4710

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Downloaded from www.industrydocuments.ucsf.edu/docs/hygl0001

health-care costs due to defendants' alleged 1965 misconduct is different in 1970 than it is in 1990. Specifically:

> In 1970, the measure of damages is a comparison between a thirty-five-year-old who has smoked continuously for 15 years in the actual world versus that same thirty-five-year-old who would have smoked for ten years but then became abstinent for five years in the counterfactual world.
>
> In 1990, the measure of damages for this same person is a comparison between a fifty-five-year-old who smoked for 30 years and has been abstinent for five years in the actual world versus that same fifty-five-year-old who would have smoked for only ten years and quit a quarter century ago in the counterfactual world.

The measure of any increased health-care costs of the individual due to the alleged misconduct by defendants toward the individual plainly will be different in different years, requiring a longitudinal analysis commencing with the date of the alleged misconduct, or its first effect, and continuing through each subsequent, relevant year.

Also, notice that if this individual's smoking behavior were unaffected by the alleged misconduct, as in counterfactual world 2, and, therefore the person would have the same smoking behavior history in the counterfactual world as in the actual world, there would be no effect of the alleged misconduct on health-care costs for the individual.

Now consider the excess health-care costs to the trust caused by the effect of the alleged misconduct on the trusts. Relative to the actual world, in both counterfactual world 1 and counterfactual world 2, there can be no excess health-care costs incurred by the trust due to the effect of the defendants' alleged misconduct on the trusts: the trust's behavior was unaffected by the alleged misconduct. That is, although there are individual-initiated changes in behavior in

Produced by R.J.R.T.C. Inv. Library

52299 4711

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Digitized by the Internet Archive http://www.industrydocuments.ucsf.edu/docs/hygl0001

counterfactual world 1 (but not in counterfactual world 2), there are no trust-initiated changes in either counterfactual world 1 or counterfactual world 2.

Now consider counterfactual worlds 1 and 3: the individual's behavior is the same in both counterfactual worlds, unaffected by the trust's modified behavior due to the alleged misconduct. In this scenario, once again, there are no excess costs to the trust due to the effect of the alleged misconduct on the trusts' behavior because the individual's behavior is unaffected by the trust's behavior; the changes in behavior from the actual world are solely individual-initiated.

Finally, consider counterfactual worlds 2 and 4. In this comparison, the increase in the trust's excess health-care costs for this individual due to the effect of the alleged misconduct on the trust are the extra costs from 1970, when the trust initiated its preventative smoking program and affected the individual's smoking behavior, through the end of the relevant period. This amount is in contrast to the potential increase in medical costs for the individual caused by the alleged misconduct, which would begin to accrue in 1965 if that individual's smoking behavior changed in 1965, as in counterfactual world 1.

2. Example Two

In this example, there is the actual world, as it exists with the defendants' alleged misconduct, and two counterfactual worlds without the defendants' alleged misconduct. In counterfactual world 1, without the alleged misconduct, there were no behavioral changes for either the individual or the trust. In counterfactual world 2, without the alleged misconduct, in contrast, in 1975 a trust-initiated smoking prevention program caused the individual to stop smoking, but the

Produced by R.J.R. Inc. in HIV litig.

52299 4712

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Stid/oeq07a00/pdf www.industrydocuments.ucsf.edu/docs/hygl0001

cessation of smoking by the individual was accompanied by overeating,[1] which had its own adverse health effects.

	Actual World with Alleged Misconduct Starting in 1965	Counterfactual Worlds: Defendants Disseminate Information on Health Risks	
		Counterfactual World 1: No Change in Trust Behavior	Counterfactual World 2: Trust Initiates Smoking Prevention Program
1935	Born; has family history of heart disease	Born; has family history of heart disease	Born; has family history of heart disease
1955	Starts smoking heavily; normal diet	Starts smoking heavily; normal diet	Starts smoking heavily; normal diet
1965	Defendants withhold information on health risks - continues smoking	Defendants disseminate information on health risks - continues smoking	Defendants disseminate information on health risks - still smoking
1975	Still smoking heavily; slightly overweight	Still smoking heavily; slightly overweight	Trust starts smoking prevention program - quits smoking; gains weight
1985	No change	No change	Becomes severely overweight
1994	Contracts lung cancer and dies	Contracts lung cancer and dies	No change
1995	No longer alive	No longer alive	Has heart attack and bypass surgery
1996	"	"	Has second heart attack.
2004	"	"	Still receiving health benefits from trust
2005	"	"	Dies

This example illustrates additional essential characteristics of the proper data collection and modeling approach beyond the need: (a) to consider the behavioral modification of trusts and

[1] A not unlikely possibility. *See*, *American J. of Epidemiology* 1998; 148:821-830, 831-832, suggesting that women who permanently quit smoking gained, on average, 19.2 pounds over five years. The corresponding weight gain for men was 16.7 pounds.

produced by IRTC in HUVEHEY

52299 4713

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf www.industrydocuments.ucsf.edu/docs/hygl0001

individuals in counterfactual worlds, (b) to consider background and other confounding characteristics, (c) to perform the analysis over time, and (d) to explicate assumptions. First, it illustrates that one must consider the possibility that smokers' other health-related behaviors (*e.g.*, overeating or alcohol consumption) might be different in a counterfactual world without the alleged misconduct. Second, the example illustrates that trust expenditures can be higher in a world without the alleged misconduct than in a world with the alleged misconduct.

In particular, in the actual world in this example, the trust bears the medical expense of this individual's lung cancer and related death in 1994, whereas in counterfactual world 2, the trust bears the medical expense of this individual's first heart attack and by-pass surgery in 1995, as well as the second heart attack in 1996 and all medical costs until the individual dies in 2005. Clearly, the costs borne by the trust caused by its own behavior in the counterfactual world will be substantially larger in that counterfactual world than either in the actual world or in counterfactual world 1 without trust-initiated effects on this individual because of the extra expenditures needed to treat the heart attacks, by-pass surgery and late-life medical expenditures.

The health-care expenditure model also would have to take into account the passage of time because costs accumulate in time. Instead of comparing total costs through the end of the damages period, however, one could have compared costs up to some event in time, such as earliest death in any world. Thus, for instance, in my example two, where the earliest death of this hypothetical individual occurs in 1994, his health-care costs stream through 1994 would be less in counterfactual world 2 without the defendants' alleged misconduct and with trust-initiated smoking prevention programs than in the actual world with defendants' alleged misconduct (or in counterfactual world 1

Produced in Litigation
HUMPHREY

52299 4714

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

without the alleged misconduct or changes in the trust's behavior) where the trust bears the cost of treating the individual's lung cancer.

Because these alternative worlds are counterfactual, there is no direct evidence to estimate how likely each world would have been in the absence of the alleged misconduct. Instead, evidence from data observed in the actual world must be coupled with assumptions to estimate the likelihood of alternative counterfactual worlds.

Finally, from a statistical perspective, such an analysis in each counterfactual world need not be done for each individual who was or would have been a participant in the plaintiff trusts. Rather, such an analysis is needed for each subgroup defined by the pertinent smoking behaviors, background characteristics and confounding factors discussed previously. The subgroup specific results are then aggregated across all the subgroups to create an estimate of each trusts' costs due to the alleged misconduct.

3. Comments On The Task

Drawing causal inferences from observational data in this way requires care. Ideally the effort would involve the input of a team of experts, including ones having knowledge about medical expenditure models and behavioral models. Nevertheless, there is a straightforward scientifically and statistically valid framework for drawing causal inferences about any medical expenditures due to the defendants' alleged misconduct.[2]

---

[2] An overview of this framework is stated in Rubin (1998) "What does it mean to estimate the causal effects of 'smoking'?" Invited paper, August 10, 1998, American Statistical Association Section on Epidemiology. More aspects of this approach were presented at an invited Plenary Talk "Estimating the Causal Effects of Smoking" for the Centers for Disease Control and Prevention, Atlanta, Georgia, January 29, 1999.

52299 4715

15 of 37

http://legacy.library.ucsf.edu/tid/oqq07a00/pdf Byttps://www.industrydocuments.ucsf.edu/docs/hygl0001

To be scientific, the underlying assumptions must be explicated in detailed and disaggregated forms. This way, the assumptions can be assessed individually for plausibility and can be altered individually to allow observation of the consequences on answers. When all the assumptions are bundled together, the resultant analysis may be little more than a subjective assumption of the answer.

Statistically, each instance of defendants' alleged misconduct (specified by its character and its timing) can be viewed as defining a level of a factor in a hypothetical factorial experiment. For example, one factor could correspond to the defendants' alleged failure to disseminate adequately information regarding the health risks of smoking, and levels of this factor could correspond to the kinds of information and their dates of dissemination. A second factor could be the alleged failure of defendants to market "safer" cigarettes. The combinations of different levels of these factors would define the defendants' conduct in a counterfactual world without specific acts of alleged misconduct. Issues of additive versus synergistic effects of the alleged acts of misconduct would be addressed by posited interactions in the hypothetical experiment. For instance, in the absence of all interactions, the costs due to specific alleged acts of misconduct would simply be the aggregate (sum) of the costs across the specific, individual alleged acts of misconduct. This framework allows the explication and disentangling of assumptions needed to address the question of the effect on the trusts' health care costs due to defendants' alleged misconduct.

Apart from any applicable legal considerations, the question of the effect of defendants' alleged misconduct on the health-care expenditures of the plaintiffs, in principle, can be addressed by statistical analyses of appropriate data coupled with explicit statistical assumptions. Any statistical analysis that can address this question must distinguish between smoking behavior affected

Privileged and Confidential Attorney Work Product

522994716

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Support provided by the UCSF Library www.industrydocuments.ucsf.edu/docs/hygl0001

and unaffected by that misconduct; take into account the passage of time; adjust reliably for differences in background characteristics and other confounding factors between those exposed to cigarette smoke and those not exposed to cigarette smoke; focus on the relevant population; explicate assumptions; use valid statistical methods; and account for the timing and nature of defendants' alleged misconduct (*e.g.*, exclude any increased health-care costs due to smoking that occurred before any alleged misconduct by defendants).

II. HOW TO ADJUST FOR BACKGROUND AND OTHER CONFOUNDING DIFFERENCES BETWEEN SMOKERS AND NONSMOKERS.

Any analysis of smoking-attributable expenditures due to the mere existence of smoking or due solely to the defendants' alleged misconduct must properly adjust for differences in covariates (background and other confounding variables) in a medical-expenditures model. All of the plaintiffs' experts in this litigation simply rely upon studies performed by other researchers for such adjustments. However, none of the studies relied upon by Dr. Dement or Dr. Harris that I have reviewed properly adjusts for important differences between smokers and nonsmokers. Instead, these studies utilize coarse categorical or regression techniques, which, based on my analysis of differences in the distribution of covariates between smokers and nonsmokers, produce results that do not reliably adjust for important covariate differences between smokers and nonsmokers. To obtain reliable adjustments for covariates, it is important to combine regression modeling with matching and/or subclassification using propensity scores.

A. Background on Adjusting for Covariates in Observational Studies

The statistical literature warns that regression analysis cannot reliably adjust for differences in covariates when there are substantial differences in the distribution of these covariates in the two groups.

Produced Pursuant to Tobacco Litigation Master Settlement Agreement

52299 4717

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf http://www.industrydocuments.ucsf.edu/docs/hygl0001

For example, William G. Cochran, who served on the Advisory Committee that wrote the 1964 Surgeon General's Report, wrote extensively on methods for the analysis of observational studies, as summarized in my chapter on his work on this topic in *W.G. Cochran's Impact on Statistics* (Rao, 1984, John Wiley, New York). In 1957 Cochran wrote:

> "... when the x-variables show real differences among groups -- the case in which adjustment is needed most -- covariance adjustments [i.e., regression adjustments] involve a greater or less degree of extrapolation. To illustrate by an extreme case, suppose that we were adjusting for differences in parents' income in a comparison of private and public school children, and that the private-school incomes ranged from $10,000-$12,000, while the public-school incomes ranged from $4,000-$6,000. The covariance would adjust results so that they allegedly applied to a mean income of $8,000 in each group, although neither group has any observations in which incomes are at or even near this level." Cochran, William G. "Analysis of Covariance: Its Nature and Uses." *Biometrics*, Vol. 13, pp. 261-281.

And in 1965, he wrote:

> "If the original x-distributions diverge widely, none of the methods [e.g. regression adjustment] can be trusted to remove all, or nearly all, the bias. This discussion brings out the importance of finding comparison groups in which the initial differences among the distributions of the disturbing variables are small."

And in the same article:

> "With several x-variables, the common practice is to compare the marginal distributions in the two groups for each x-variable separately. The above argument makes it clear, however, that if the form of the regression of y on the x's is unknown, identity of the whole multi-variate distribution is required for freedom from bias." Cochran, William G. "The planning of observational studies of human populations." *The Journal of the Royal Statistical Society, A*, Vol. 128, pp. 234-265.

In particular, there are three basic distributional conditions that in general practice must simultaneously obtain for regression adjustment (whether by ordinary linear regression, linear

52299 4718

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

logistic regression or linear-log regression) to be trustworthy. If any of these conditions is not satisfied, the differences between the distributions of covariates in the two groups must be regarded as substantial, and regression adjustment is unreliable and cannot be trusted. These conditions are:

1. The difference in the means of the propensity scores in the two groups being compared must be small (*e.g.*, the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.

2. The ratio of the variances of the propensity score in the two groups must be close to one (*e.g.*, ½ or 2 are far too extreme).

3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (*e.g.*, ½ or 2 are far too extreme).

Specific tabulations and calculations relevant to these points can be found, for example, in Cochran and Rubin (1973) "Controlling Bias in Observational Studies: A Review," *Sankhya*, Series A, Vol. 35, Part 4, pp. 417-446; Rubin (1973) "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies," *Biometrics*, 29, pp. 185-203; and Rubin (1979) "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, pp. 318-328.

In particular, Cochran and Rubin (1973, p. 426) state that "linear regression on random samples gives wildly erratic results . . ., sometimes markedly overcorrecting or even (with B = 1/4 for e^X) greatly increasing the original bias [when the ratio of the variances is one-half]". Table 3.2.2 in that article implies that, when the ratio of the variances of any covariate is one half, regression can

52299 4719

19 of 37

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Generated by www.industrydocuments.ucsf.edu/docs/hygl0001

grossly overcorrect for bias or grossly undercorrect for bias. Relevant results from that table are summarized in Table 1 of my July 1998 Oklahoma Report, attached as Exhibit 1 to this report. These three guidelines and Table 1 in that report also address regression adjustments on the logit or linear log scale because they, too, rely on linear additive effects in the covariates (for discussion of this point, *see e.g.*, Anderson, Auquier, Hauck, Oakes, Vandaele, and Weisberg (1980), *Statistical Methods for Comparative Studies*, John Wiley, New York, p. 164).

B. Propensity Score Analyses Comparing Smokers and Never Smokers In The National Population As Represented By NMES and NHIS

To obtain an indication of how different smokers and never smokers might be in the national population, I first conducted two propensity score analyses using NMES (the National Medical Expenditure Survey), as discussed in my January 1998 Report in Minnesota[3] and in my July 1998 Report in Oklahoma. Those analyses, however, were unweighted to mirror what the plaintiffs' experts had done. In Table 1 below and in the electronic media I have produced in this case, I compared smokers and nonsmokers by conducting a propensity score analysis using NMES and its weights, thereby representing the national population corresponding to the NMES design. In that analysis, I used the 32 covariates selected by the plaintiff's expert, Dr. Harrison, in his Oklahoma report.

Produced in Litigation
HUNDLEY

52299 4720

[3] The January 1998 Minnesota Report is attached here as Exhibit 2; the July 1998 Oklahoma Report is attached as Exhibit 3.

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Document hosted at http://www.industrydocuments.ucsf.edu/docs/hygl0001

Table 1: Propensity Score Analyses For Smokers Versus Never Smokers In NMES and NHIS

Smoking Group	Propensity Score		Residuals Orthogonal to Propensity Score				
			Variance Ratios in Range				
	Bias	Variance Ratio	(0, ½)	(½, 4/5)	(4/5, 5/4)	(5/4, 2)	(2, ∞)
Current (NMES)	.81	.97	0	6	22	3	1
Former (NMES)	.75	1.24	0	5	21	4	2
Current Females (NHIS)	.62	.41	15	14	61	30	15
Current Males (NHIS)	.89	.45	21	23	50	37	4
Former Females (NHIS)	.50	.30	13	25	73	20	4
Former Males (NHIS)	.87	.30	12	35	67	10	11

I also conducted four propensity score analyses using the 1991 NHIS (the National Health Interview Survey) data, using sampling weights and all of the covariates selected by the plaintiff's experts, the Cambridge Team, in the Massachusetts litigation,[4] separately by gender, as in the Cambridge Team's analysis. Those results are also in Table 1.

[4] Drs. Cutler, Epstein, Frank, Hartman, King, and Newhouse, "The Impact of Smoking on Medicaid Spending in Massachusetts 1970-1997," June 15, 1998.

52299 4721

21 of 37

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf

Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

In all instances, the differences between smokers and nonsmokers nationally are substantial enough that regression models cannot reliably adjust for even the background characteristics and other confounding factors included in the regressions.

Thus, the linear regression methods used by the studies of the national population cited by Dr. Harris cannot be considered as having adjusted in a reliable way for differences between smokers and nonsmokers. Better procedures would combine regression methods and propensity score subclassification, as suggested by the statistical literature.

C. Propensity Score Analyses Comparing Smokers and Never Smokers In the Union Subset of NMES and NHIS

The differences between smokers and nonsmokers observed nationally appear to be even more substantial in the union subsets of the NMES[5] and NHIS[6] data. These results using the sampling weights are presented in Table 2. Even if the experts for the plaintiff union trusts had examined a union population or relied upon studies in the literature that examined a union population, they must make reliable adjustments for differences between smokers and nonsmokers on background characteristics and other confounding factors; linear regression adjustments are not reliable in this setting.

---

[5] I specifically looked at those individuals who reported that they had private insurance provided by a union in any of the four rounds of interviews. To compute the propensity scores, I again used the 32 variables in the Harrison Report in Oklahoma. Additional detail is found in electronic materials accompanying my reports in this case.

[6] I examined in NHIS the same population that Dr. Dement used in his expert report in the Ohio union litigation as a surrogate for the union population when he estimated smoking prevalence values. To compute the propensity scores, I used 34 of the variables employed by the Cambridge Team's June 15, 1998 Report in Massachusetts, excluding industry and occupation variables, and collapsing some other indicator variables; additional detail is found in the electronic materials accompanying my reports in this case.

52299 4722

22 of 37

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf

Produced by the HUVARD Document Depository http://www.industrydocuments.ucsf.edu/docs/hygl0001

Table 2: Propensity Score Analyses For Smokers Versus Nonsmokers In Union Subsets of NMES and NHIS

Smoking Group	Propensity Score		Residuals Orthogonal to Propensity Score				
	Bias	Variance Ratio	Variance Ratios in Range				
			(0, ½)	(½, 4/5)	(4/5, 5/4)	(5/4, 2)	(2, ∞)
Current (NMES)	.59	.11	4	5	16	5	2
Former (NMES)	.69	.23	5	7	16	4	0
Current (NHIS)	.86	.76	2	5	24	3	0
Former (NHIS)	.95	.96	0	6	20	5	3

D. More Reliable Adjustments Combine Regression Modeling and Propensity Score Methods

There is a quarter century of statistical literature, (*e.g.*, Cochran and Rubin, 1975) suggesting that regression modeling alone is unreliable in the situations presented. This same literature suggests, however, that combining regression modeling with matching and/or subclassification using the propensity score is substantially more reliable. For example, in addition to the references cited above, see Roseman (1998) and Rubin and Thomas (1999). Moreover, based upon my review of a literature search, propensity score methods are now being relatively widely referenced in highly respected medical and statistical journals.

Produced by HUVER LIBRARY

52299 4723

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Byttps://www.industrydocuments.ucsf.edu/docs/hygl0001

III. DR. HARRIS

A. General Comments

Dr. Harris makes no effort to estimate separately the effect of defendants' alleged misconduct on the behavior of the plaintiff trusts and, in turn, the effect of that change in trust-initiated smoking behavior on the trusts' health-care expenditures. Dr. Harris does present estimates of the effect of the defendants' alleged misconduct generally on the trusts' health-care expenditures. His analyses, however, have essentially none of the characteristics that are necessary to estimate the health-care costs incurred by the trusts because of the existence of smoking or defendants' alleged misconduct generally. As detailed more fully below, Dr. Harris' analyses fail to a) take into account the passage of time in the manner I described; b) control properly for important background and other confounding characteristics; c) distinguish between different types of smoking behaviors that can lead to different health-care expenditure outcomes; d) focus on the participants in the plaintiff trusts; e) disaggregate assumptions in a way that enables the assumptions to be assessed and varied individually to see the consequences on the resulting estimates f) consider the possibility that smokers' other health-related behaviors may also change in the counterfactual world; or g) employ appropriate statistical methods or rely on the results of studies that do so.

The way Dr. Harris' analyses attempt to estimate the effect of defendants' alleged misconduct generally on the trusts' health-care expenditures is as follows. He first attempts to estimate the proportion of smoking-attributable expenses that are due to the alleged misconduct. This estimation has three parts: (1) general estimation of medical expenditures due to the existence of smoking based on his review of selected literature; (2) his general subjective estimation of medical expenditures that would have existed in a counterfactual world without the misconduct; and (3) the

Produced by: HIGHLIGHT PRINTING

52299 4724

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf Stanton Glantz http://www.industrydocuments.ucsf.edu/docs/hygl0001

calculation of the resulting proportionate differences, these being his estimated proportions of estimated smoking-attributable expenses due to the alleged misconduct. Dr. Harris then applies these estimated proportions to the estimated smoking-attributable expenses provided by Mr. Roberts using the mortality-based smoking-attributable fractions estimated by Dr. Dement. All of the analyses conducted by and relied on by Dr. Harris that I have examined are unreliable and statistically invalid for this purpose.

B. The Existence of Smoking

To estimate the health-care costs incurred by the plaintiff trusts as a result of the existence of cigarette smoking (as opposed to defendants' alleged misconduct), one has to compare plaintiff trusts' health-care costs in the actual world to what those costs would have been in a counterfactual world in which no cigarette smoking existed. To be reliable and valid, such an estimate must possess many of the same essential characteristics that I described above in Section I of this report. Dr. Harris' estimates do not have those characteristics. His estimates of these smoking-attributable expenditures depend on two quantities: the relative expenditure ratios of ever smokers to never smokers, or "r," and the prevalence of ever smoking, or "p." His actual world values for both "p" and "r" are unreliable.

Dr. Harris' actual world ever smoking "p" values are based on a piece-wise linear model without any provided statistical or economic justification. The "knot" value, which Dr. Harris places at 1974, is not calculated from any particular data set, nor allowed to vary by background characteristics and other confounding factors of individuals, but is simply postulated by Dr. Harris, based on his subjective assessments. The line reflecting ever smoking prevalence from 1953 to 1974 does not appear to be a close match to the data points on which Dr. Harris relies, and there is no

Produced by R.J. Reynolds Tobacco Company in HUNDLEY

52299 4725

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf
Additional support provided by www.industrydocuments.ucsf.edu/docs/hygl0001

confidence interval for his determinations. even when they appear to be based on the analysis of data under particular assumptions.

In determining his values for "r," Dr. Harris relies on a group of studies of health-care expenditures. There are a variety of problems with those studies, illustrated below.

First, it is well known in many branches of statistics, epidemiology, and economics that it is necessary to control for background and other confounding variables when trying to estimate causal effects. These variables define subpopulations where, for example, similar smokers and nonsmokers can be realistically compared with respect to health outcomes, including health-care expenditures. Even when interest focuses solely on aggregate costs and not on subpopulations, background and other confounding variables that define subpopulations must be considered to obtain valid estimates of aggregate costs.

Nevertheless, some of the studies relied on by Dr. Harris do not even attempt to control for possibly important confounding variables; others attempt to do so but do not adjust for those factors reliably. My review suggests that none of the studies relied on by Dr. Harris for his values of "r" adjust for the full set of background characteristics that even the experts retained by the plaintiffs in similar cases have adjusted for. *See, e.g.*, L.S. Miller, X. Zhang, T. Novotny, D. P. Rice, and W. Max, "State Estimates of Medicaid Expenditures Attributable to Cigarette Smoking, Fiscal Year 1993," *Public Health Reports* 113:140-151 (March/April 1998) at 141-42; *Minnesota* Trial Testimony of Dr. Samet at 3482, Deposition Testimony of Dr. Samet at 369-70, and Trial Testimony of Dr. Wyant at 5372-73 and 5391-93; *Oklahoma* April 27, 1998 Report of Dr. G. W. Harrison at 18-20 and 39; Vincent Miller, *et al.*, *Smoking Attributable Medicaid Care Costs: Models and*

52299 4726

26 of 37

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Byttps://www.industrydocuments.ucsf.edu/docs/hygl0001

Results, September 3, 1997, Tables 2.4-2.8 at 23-27 and 29-30; David M. Cutler, *et. al.*, *The Impact of Smoking On Medicaid Spending In Massachusetts: 1970-1997*, June 15, 1998, Table IV.2.[7]

Moreover, none of these studies relied upon by Dr. Harris that I have examined reliably adjusts even for the background characteristics they purport to adjust for. For example, as documented in my January 1998 Supplemental Report in Minnesota, in my July 1998 Report in Oklahoma, my November 20, 1998 Supplemental Report in the Ohio Union litigation, and in Section II in this report, the differences between smokers and nonsmokers in NMES and NHIS are so substantial that the attempted adjustments using regression models are known to be entirely unreliable. Furthermore, such differences persist even within the union-worker subset of NMES and NHIS as shown in Section II of this report.

Second, none of the studies on which Dr. Harris relies explicitly considers a union trust fund recipient population, much less a population of union trust fund participants in the Northwest United States. Dr. Harris implicitly appears to assume that relative expenditure ratios of ever smokers compared to never smokers computed based on the general U.S. population or populations of employees of a given firm apply equally to the participants in the plaintiff trusts, even though the trust participants may differ in important ways from these other populations. This is despite the fact that in deposition, Dr. Harris allows, as an explanation for empirical results that counter one of his theories, that Ohio union members may differ in important ways from national union members. *See* Harris Deposition (Ohio), December 15, 1998 at 86-89. Similarly, the ratio of medical expenditures between nonsmokers and smokers for the participants in the plaintiff trusts may differ from that of

[7] Ideally, any characteristic that might be considered important should be used because a subsequent analysis can always average over that characteristic, thereby ignoring it, whereas an analysis that leaves it out initially has that characteristic confounded with smoking behavior.

Produced in LITIGATION
Privileged

52299 4727

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Additional digitized by www.industrydocuments.ucsf.edu/docs/hygl0001

the national general population or of employees of a given firm. Yet, Dr. Harris' analysis provides no evaluation of whether the values of "r" he uses are applicable to the participants in the plaintiff union trusts. Section IV of this report documents the substantial differences between union workers and the rest of the national population as represented in NMES and in NHIS.

Third, based on my review, many of these studies relied on by Dr. Harris confront missing data problems, but, as I have described in my reports in other cases (see Exhibits 2 and 3), they use techniques documented in the statistical literature as being flawed. *See, e.g.*, Little and Rubin, *Statistical Analysis with Missing Data*, (John Wiley, NY 1987); Panel on Incomplete Data, National Academy of Sciences, *Incomplete Data in Sample Surveys*, Vol. 1-3 (Academic Press, NY 1983).

For example, many of the studies relied on by Dr. Harris impute values for the data that are missing and use the resulting data sets as if the imputed values were real data. But, in the instances I have examined, the studies do not use sound statistical methods to impute these missing data, and, instead, use scientifically invalid procedures. Among the errors made by the studies relied on by Dr. Harris for his values of "r" are: using best predicted values for missing data, imputing the missing data sequentially without proper conditioning, imputing the missing data using only part of the available known data, imputing the missing data without taking into account their uncertainty, relying on strong unsubstantiated assumptions, using arbitrary values to replace missing values, and failing to conduct any informative sensitivity analyses of the imputations, such as varying fundamental assumptions underlying the imputations. As I have described in detail in other reports, these kinds of errors mean that the data analyses and conclusions about "r" values relied on by Dr. Harris are statistically unreliable and invalid.

Confidential Pursuant to Protective Order

52299 4728

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Additional digitized by www.industrydocuments.ucsf.edu/docs/hygl0001

Fourth, Dr. Harris' analyses of both "p" and "r" merely compare never and ever smokers (as a weighted average of former and current smokers). Differences in smoking behaviors that epidemiologists, physicians, and even Dr. Harris,[8] appear to believe are important – *e.g.*, smoking intensity, smoking duration, type of cigarette smoked, or time since quitting – are nowhere accounted for in his analyses of either "p" or "r." To be valid, such differences, as discussed above, must be taken into account in appropriate subgroups defined by demographic characteristics and other health-related behaviors.

Finally, Dr. Harris' analyses do not take into account the passage of time, consider the possibility that other health-related behaviors (*e.g.*, overeating or illicit drug behavior) may be affected by the existence of smoking, or explicate necessary assumptions in a manner that allows them individually to be identified and evaluated or altered to determine the effects on the results.

C. The Effect of Defendants' Alleged Misconduct Generally

Dr. Harris does not attempt to assess separately the effect of defendants' alleged misconduct on the trusts' behavior specifically (two-component behavioral model).

In his December 30, 1998 report in this case, Dr. Harris does attempt to estimate the effect of defendants' alleged misconduct generally on the trusts' health-care expenditures (one-component behavioral model). He presents results from his one-component analysis. Although Dr. Harris properly recognizes the need to consider separately the effect of defendants' alleged misconduct on smoking behavior and the effect of that change in smoking behavior on the trusts' expenditures, the estimates he presents are neither reliable nor statistically valid.

Produced by RJR
in HUMPHREY

[8] *See, e.g.*, Harris Oct. 26, 1998, "Notes on Changes Over Time in the Mean Duration of Quitting Among Former Smokers: Impact on Θ, the Reduction in the Excess Risk of Quitting." *See also*, Undated original report of Dr. Lauer in the Ohio Union litigation, entitled "Smoking and Coronary Health Disease" at 12-14, and his supplemental October 22, 1998 report, "Smoking and Stroke" at 4.

52299 4729

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Dr. Harris' estimates, as an initial matter, critically depend upon his estimates of the actual world values of "p," and "r," and his resulting estimates of the smoking-attributable fraction of health-care costs due to the very existence of smoking. Because those estimates, as discussed above, are unreliable and invalid, so, too, are Dr. Harris' estimates of the effect of defendants' alleged misconduct, even if his assessment of the change in smoking behavior of the trusts' participants and the consequent change in the health-care expenditures of the trusts in the counterfactual world were reliable and valid.

Dr. Harris' further analyses of the effect of the defendants' alleged misconduct on smoking behavior and health-care costs in the counterfactual world, additionally, are neither reliable nor statistically valid. Those analyses have essentially none of the characteristics that are necessary to estimate the health-care costs incurred by the trusts because of defendants' alleged misconduct. As with his actual world estimates, Dr. Harris fails to: a) take into account the passage of time in the manner I described; b) control for important background and other confounding characteristics; c) distinguish between different types of smoking behaviors that can lead to different health-care expenditure outcomes; d) focus on the participants in the plaintiff trusts; e) disaggregate assumptions in a way that enables the assumptions to be assessed and varied individually to see the consequences on the resulting estimates; f) consider the possibility that smokers' other health-related behaviors may also change in the counterfactual world; or g) employ appropriate statistical methods.

In the end, Dr. Harris' consideration of the effect of defendants' alleged misconduct is really a subjective statement based on his opinions of what would have happened in a counterfactual world without that alleged misconduct. Two illustrations help demonstrate the point.

Produced by the Tobacco Industry Archive

52299 4730

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf http://www.industrydocuments.ucsf.edu/docs/hygl0001

First, Dr. Harris appears to conclude that, in the absence of the alleged misconduct, the tobacco companies would have introduced "safer" products to the marketplace. By this he appears to mean both that products that never were introduced would have been introduced and would have been adopted by consumers, and that the "safer" products factually introduced would have been introduced sooner. But in his report in this case, Dr. Harris never specifies: how much "safer" any "safer" product would have been than products already on the market at that time; the rate at which consumers, and participants in the plaintiff trusts in particular, would have adopted them; or what the effects of switching, if any, to any of these "safer" cigarettes would have been on the trusts' health-care expenditures over time. Nor could he do so on deposition. *See, e.g.*, Harris Washington Dep. At 396, 398, and 403; Harris Arizona Dep. At 18, 20, 297, 322-25, and 346-52.

Second, Dr. Harris generally concludes that, in a counterfactual world without the alleged misconduct, the tobacco industry would have disseminated information about the health risks of smoking and this information would have modified the behavior of those who were smokers in the actual world. It appears, however, that many factors affect consumers' decisions to begin to smoke, to quit smoking, or to switch to "safer" cigarettes. According to the Surgeon General, smoking initiation is affected by parental smoking, peer influence, and a variety of other psycho-social risk factors. U.S. DHSS, *Preventing Tobacco Use Among Young People, A Report of the Surgeon General*, Ch. 4 (1994); *see also* Harris Washington Dep. at 25-36. Similarly, a number of factors reportedly affect consumer choices among different cigarette brands that, by some measures, pose different levels of health risks, as well as decisions by existing smokers to continue or to quit smoking. *See, e.g.*, Harris Washington Dep. 36-45; U.S. DHHS, *The Health Consequences of Smoking, The Changing Cigarette, a Report of The Surgeon General* (1981) at 5-6; Redmond, W.H.,

PRIVILEGED AND CONFIDENTIAL ATTORNEY WORK PRODUCT

52299 4731

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Support provided by the National Institutes of Health
www.industrydocuments.ucsf.edu/docs/hygl0001

"Product Disadoption: Quitting Smoking as a Diffusion Process," *J. of Pub. Policy & Marketing* 15(1):87-97 (Spring 1996) at 87 *citing* U.S. DHHS. *Reducing the Health Consequences of Smoking: 25 Years of Progress* (1989). Yet, Dr. Harris incorporates no methodology to quantify the role those factors play in consumer choices. He consequently cannot reliably assert the role that the industry's alleged misconduct played in consumer choices.

For instance, Dr. Harris obtains his counterfactual world values for "p" by shifting back his "knot" value or peak for ever-smoking prevalence from his factual world estimate of 1974 to 1955, 1960, or 1965, without any empirical justification whatsoever. *See, e.g.*, Harris Washington Dep. at 456-457.

IV. ESTIMATES OF PREVALENCE OR RELATIVE RISK BASED ON THE NATIONAL POPULATION ARE UNRELIABLE ESTIMATES FOR THE UNION POPULATION.

Dr. Harris relies upon estimates of relative risk from the literature that were derived using national data or data for particular firms, and he, moreover, estimates smoking prevalence values based on national data. Thus, he implicitly assumes a) that the quantitative impact of smoking on health-care expenditures is the same in the national population or for employees of selected firms as it is among individuals who participate in health-care programs funded by union trust funds; and b) that smoking prevalence estimates using national data apply equally to the participants in the plaintiff union trusts.

Despite Dr. Harris' deposition testimony that "... in an ideal world, ... one would really want to look specifically at the population at hand ...", nowhere does Dr. Harris assess the validity of these assumptions. Harris Maryland Dep. at 66-67 (discussing the African-American population). I have prepared propensity score analyses comparing the union populations represented in the NMES

52299 4732

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

and NHIS samples to the rest of the population as represented in those samples, using their sampling weights. The results are in Table 3 using the 43 variables in the Harrison Report in Oklahoma for NMES, and using 40 of the covariates plus two indicators for current and former smoking from the Cambridge Team's June 15 Massachusetts report for NHIS, excluding industry and occupation indicators.[9] In both instances, the analyses revealed that there are substantial differences between the union population and the rest of the nation on the background characteristics and other confounding factors that other plaintiffs' experts in litigation against these defendants believe are important. The differences are so substantial that estimates of relative expenditure risk ratios or smoking prevalence values based on national data cannot reliably be applied to the union population nationally, much less to the participants of the plaintiff Northwest Laborers union trusts.

Dr. Dement's estimate of the relative mortality risk of smokers for particular diseases likewise is based on the American Cancer Society's Cancer Prevention Survey II dataset, which is not representative of either the national population or the union population in the United States. I plan to conduct propensity score analyses to reveal the size of the differences between these populations and the ACS CPS II dataset.

Moreover, it has been documented that smoking trends in the blue-collar population are different than in the rest of the population, typically declining more slowly in time than in the general population. *See, e.g.*, Nelson, *et al.*, "Cigarette Smoking Prevalence by Occupation in the United States," J. Occup. Med., Vol. 36, No. 5, May 1994, pp. 516-25.

[9] The variables exclude occupation and industry indicators, and collapse some other indicators into coarser categories; a complete list of the variables used in these propensity score analyses is found in the electronic materials accompanying this report.

Produced by the Tobacco Products Liquidating Trust

52299 4733

33 of 37

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf

Downloaded from http://www.industrydocuments.ucsf.edu/docs/hygl0001

Table 3: Propensity Score Analyses For Union Versus Non-Union Populations as Represented in NMES and NHIS

Data Source	Propensity Score		Residuals Orthogonal to Propensity Score				
	Bias	Variance Ratio	Variance Ratios in Range				
			(0, ½)	(½, 4/5)	(4/5, 5/4)	(5/4, 2)	(2, ∞)
NMES	.80	.81	4	4	30	4	1
NHIS	1.19	.31	7	5	23	5	0

DR. DEMENT AND MR. ROBERTS

Neither Dr. Dement nor Mr. Roberts attempts to quantify the impact of defendants' alleged misconduct on the trusts' health-care expenditures at all, either generally (one-component behavioral model) or specifically due to changes in the trusts' behavior (two-component behavioral model).

Instead, they exclusively attempt to address a casual question different from one involving defendants' alleged misconduct: What health-care expenditures were incurred by the plaintiff trusts as a result of the existence of cigarette smoking, regardless of defendants' alleged wrongdoing? As a result of attempting to address this other question, their estimates of smoking-attributable expenditures, even if all of the errors in their underlying analyses were eliminated, would still substantially overstate the health-care expenditures due to defendants' alleged wrongful conduct. This fact is reflected in Dr. Harris' allocating only a portion of the dollar amounts from the Dement and Roberts reports to the alleged misconduct.

The analyses of Dr. Dement and Mr. Roberts, moreover, fail to estimate reliably and validly the health-care costs incurred by the plaintiff trusts due to the existence of smoking because those analyses do not have the essential characteristics that I described above in the first section of this report as necessary to generate reliable and statistically valid results:

52299 4734

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Support provided by the Industry Documents Library: http://www.industrydocuments.ucsf.edu/docs/hygl0001

1. Dr. Dement and Mr. Roberts do not even model medical expenditures, but rather disease mortality.

2. Their analyses do not attempt to control adequately for background and other confounding variables, other than gender and, in some instances, broad age groups.

3. Their reports distinguish only between current, former, and never smoking, disregarding differences in these categories of smoking behavior that physicians and epidemiologists believe are important.

4. Their calculations fail to address the passage of time in the manner I described above.

5. Their calculations depend upon relative mortality estimates and smoking prevalence estimates that are derived from samples that are not representative of the populations covered by the trusts.

6. Their reports leave critical assumptions implicit and not evaluated. For example, nowhere in their reports do Dement and Roberts scientifically evaluate their assumption that the relative risk for mortality is the same as the relative risk for expenditures.

7. None of their calculations consider the possibility that other health-related behaviors (*e.g.*, overeating, illicit drug behavior) may be affected by the existence of smoking.

8. Missing data problems are severe but dealt with in ad hoc and inappropriate ways. For at least a dozen years (*see, e.g.*, Little and Rubin, 1987), there have been principled statistical approaches to missing data, but none of these are used or even referenced. For example:

   a. Missing ICD-9 codes (unspecified claim records) are effectively imputed using a "best prediction" scheme based at most only on coarse age, gender and status (participant/dependent), and moreover, no account is made for the uncertainty when applying the resultant composite SAFs to medical expenditures.

   b. The vast majority of the medical claims data are missing and are effectively imputed using an entirely ad hoc methodology. There are only four named but 57 absent class members, and the only "claims" data from the absent members are from 5500 records, which do not report medical expenditures. Moreover, even for the four named members, most of the data are missing for

52299 4735

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Document hosted at www.industrydocuments.ucsf.edu/docs/hygl0001

the time period under consideration. Not only are "predictions" created using at most only coarse age, gender, status and disease type, but nowhere is there any reflection of the vast uncertainty in the resulting estimated expenditures.

c. Multiple imputation (Rubin, 1987) is a principled and practical way to address missing data in many contexts. An effort to create a multiply-imputed version of NMES is proceeding under my direction. When completed, it should be able to support valid statistical inferences for estimated expenditure models.

VI. CONCLUSION

The plaintiffs' experts' analyses do not even attempt to generate any estimates of the trusts' excess expenditures that were caused by the defendants' alleged misconduct on the behavior of the trusts (two-component behavioral model).

Neither Dr. Dement nor Mr. Roberts even attempt to generate any estimates of the trusts' excess expenditures that were caused by the defendants' alleged misconduct generally (one-component behavioral model). Dr. Harris properly recognizes the need to consider separately the effect of defendants' alleged misconduct on smoking behavior and the effect of that change in smoking behavior on the trusts' health-care costs. He attempts to estimate the effect of defendants' alleged misconduct generally on the trusts' expenditures, but his estimates are neither reliable nor statistically valid. Harris' estimation is based on first estimating the proportions of health-care expenditures from the existence of smoking that are due to the alleged misconduct of the tobacco industry and, second, applying these proportions to the expenditure amounts provided by Dement and Roberts.

But the Dement and Roberts' analyses, which attempt to estimate the medical expenditures of the individuals in the trusts due to the mere existence of smoking, incorporate essentially none

Reproduced by RJRTC
HUMPHREY

52299 4736

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

of the characteristics that are critical to obtaining scientific, reliable, and statistically valid estimates and so are neither reliable nor statistically valid.

I am continuing to address the issues surrounding the plaintiffs' expert reports, and therefore may supplement my report in the future.

Donald B. Rubin, Ph.D.

Date: February 23, 1999

Produced by R.J.R.T.C.
in
TIMPEL

52299 4737

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

5/26/99

Rubin
EXHIBIT NO. 2
2-2-99
J. WILLIAMS

SECOND SUPPLEMENTAL REPORT OF DONALD B. RUBIN
NORTHWEST LABORERS

My qualifications and experience were detailed in my two prior reports in this case: my September 16, 1998 initial report and my February 23, 1999 first supplemental report.

None of the data sets upon which Dr. Dement, Mr. Roberts, and Dr. Harris rely is from a union trust fund recipient population, much less a population representative of the participants in the plaintiff Washington State union trust funds.

My prior reports used propensity score analyses to assess the differences in the distributions of pertinent background characteristics (a) between smokers and nonsmokers within the data sets relied upon by plaintiffs' experts; and (b) between the individuals represented in those data sets and a population of national union health care fund participants (represented in NMES) or a sub-population that plaintiffs' experts believe is a surrogate for a national population of union health-care fund recipients (represented in NHIS). This report extends those analyses.

Specifically, plaintiffs' experts implicitly appear to assume that the relative mortality risks, and, by further assumption, the relative expenditure ratios, of smokers compared to non-smokers, will be the same in the American Cancer Society's Cancer Prevention Survey II database ("ACS CPS-II") as they are among the participants in the plaintiff trust funds. I have now conducted propensity score analyses using a few variables comparing the ACS CPS-II data base to both the national population as a whole and within certain sub-populations defined by smoking status as represented by the NMES and NHIS probability surveys. The results of those propensity score analyses are summarized in Table 1; the computer code that generates those results together with the output are contained in the electronic materials accompanying this report.

52299 4738

1 of 7

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf
Source: http://www.industrydocuments.ucsf.edu/docs/hygl0001

Although I used only a few variables, the ACS CPS-II data are not representative of the nation, even in sub-populations defined by smoking status, such as male never smokers. There are substantial differences between the ACS CPS-II data set and the national samples: Typically, there is a bias of well over a standard deviation along the propensity score (bias > 1), with substantially less variability (both along the propensity score and in the residuals) in the ACS CPS-II data set than in the nation. That is, ACS CPS-II is not representative of the nation as a whole and is much more homogenous than the nation.

Additionally, I have included in Exhibit 1 to this report some further publications in the current epidemiological and medical literature that refer to propensity score methods. Propensity score methods are becoming increasingly popular in medicine, epidemiology, and other applied fields, such as economics, not only for assessing biases between data sets, as in Table 1, but also for obtaining more reliable adjustments between groups (such as smokers and non-smokers).

Finally, as I advised in my first supplemental report, an effort to create a multiply-imputed version of NMES to address its missing data is proceeding under my direction. Due to an unexpected and unanticipated coding error, the multiply-imputed NMES data are not yet available. That work is proceeding, however, and, provided no additional unforeseen difficulties arise, I anticipate that multiply-imputed NMES data will be available in the next few weeks.

Produced in Litigation

52299 4739

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf Source: http://www.industrydocuments.ucsf.edu/docs/hygl0001

I may also further supplement this report as appropriate to respond to the expected supplemental submissions of plaintiffs' experts.

Donald B. Rubin

Dated: May 26, 1999

produced in HUMPHREY

52299 4740

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf Downloaded from www.industrydocuments.ucsf.edu/docs/hygl0001

Table 1: Propensity Analyses Comparing Subjects in CPS-II with National Representative Samples NMES, NHIS-80, NHIS-82, and NHIS-91: Overall and by Smoking Status

Comparison Group	Propensity Score		Residual orthogonal to propensity score				
	Bias	Variance	(0, 1/2)	(1/2, 4/5)	(4/5, 5/4)	(5/4, 2)	(2, ∞)
Males							
CPS-II vs NMES	1.30	0.32	4	3	4	2	0
CPS-II vs NHIS-80	1.39	0.35	1	4	2	2	0
CPS-II vs NHIS-82	1.35	0.39	3	3	3	1	0
CPS-II vs NHIS-91	1.34	0.41	4	4	1	5	0
Females							
CPS-II vs NMES	1.10	0.48	1	5	6	1	0
CPS-II vs NHIS-80	1.26	0.51	1	4	3	3	0
CPS-II vs NHIS-82	1.19	0.56	1	3	4	2	0
CPS-II vs NHIS-91	1.10	0.66	1	2	7	4	0
Males, Current Smoker							
CPS-II vs NMES	1.46	0.48	2	5	1	3	0
CPS-II vs NHIS-80	1.44	0.46	2	4	1	2	0
CPS-II vs NHIS-91	1.53	0.51	4	3	1	2	2
Males, Former Smoker							
CPS-II vs NMES	0.92	0.42	4	3	3	1	0
CPS-II vs NHIS-80	1.06	0.51	4	2	2	1	0
CPS-II vs NHIS-91	0.95	0.54	5	2	1	4	0
Males, Never Smoker							
CPS-II vs NMES	1.53	0.27	3	4	2	2	0
CPS-II vs NHIS-80	1.56	0.27	3	3	2	1	0
CPS-II vs NHIS-91	1.48	0.43	4	4	1	2	1
Females, Current Smoker							
CPS-II vs NMES	1.29	0.53	0	6	3	2	0
CPS-II vs NHIS-80	1.32	0.57	0	4	3	2	0
CPS-II vs NHIS-91	1.32	0.67	0	5	3	4	0
Females, Former Smoker							
CPS-II vs NMES	0.98	0.61	1	7	2	1	0
CPS-II vs NHIS-80	1.16	0.58	1	4	2	2	0
CPS-II vs NHIS-91	0.86	0.92	0	4	5	3	0
Females, Never Smoker							
CPS-II vs NMES	1.07	0.43	1	3	7	0	0
CPS-II vs NHIS-80	1.20	0.48	1	2	5	1	0
CPS-II vs NHIS-91	1.11	0.68	1	3	5	3	0

Variables Used in this Analysis:
Age, Race, Marital status, Education, Body Mass Index, High blood pressure,* and Diabetes*

*Not used in NHIS-80 comparison

52299 4741

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

52299 4742

5 of 5

EXHIBIT 1

produced by RJRTC

in

HOUPHL

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

Some Recent Articles in Medical and Epidemiological Journals That Refer to Propensity Score Methods

Aitken, L., Smith, H. and Lake, E. 'Lower Medicare mortality among a set of hospitals known for good nursing care', *Medical Care*, 32, 771-787 (1994).

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C. et alia. 'The effectiveness of right heart catheterization in the initial care of critically ill patients', *Journal of the American Medical Association*, 276, 889-897 (1996).

Cook, E. F. and Goldman, L. 'Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies', *American Journal of Epidemiology*, 127, 626-639 (1988).

Cook, E. F. and Goldman, L. 'Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score', *Journal of Clinical Epidemiology*, 42, 317-324 (1989).

Drake, C. and Fisher, L. 'Prognostic models and the propensity score', *International Journal of Epidemiology*, 24, 183-187 (1995).

Eastwood, E. A. and Fisher, G. A. 'Skills acquisition among matched samples of institutionalized and community-based persons with mental-retardation', *American Journal on Mental Retardation*, 93, 75-83 (1988).

Fiebach, N. H. et alia. 'Outcomes in patients with myocardial-infarction who are initially admitted to stepdown units - data from the multicenter chest pain study', *American Journal of Medicine*, 89, 15-20 (1990).

Harrell, F. E. et alia. 'Statistical-methods in SUPPORT', *Journal of Clinical Epidemiology*, 43, S89-S98 (1990).

Kane, R. et alia. 'Improving primary care in nursing homes', *Journal of the American Geriatric Society*, 39, 359-367 (1991).

Kramer, A. M., Steiner, J. F., Schlenker, R. E., Eilertsen, T. B. et alia. 'Outcomes and costs after hip fracture and stroke', *Journal of the American Medical Association*, 277, 396-404 (1997).

Lavori, P. W. and Keller, M. B. 'Improving the aggregate performance of psychiatric methods when not all subjects receive the standard test', *Statistics in Medicine*, 7, 723-737 (1988).

Lavori, P. W., Keller, M. B. and Endicott, J. 'Improving the validity of Rh-Rdc diagnosis of major affective-disorder in uninterviewed relatives in family studies - A model based approach', *Journal of Psychiatric Research*, 22, 249-259 (1988).

DRAFT CONFIDENTIAL

52299 4743

6 of 7

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf

Batches: http://www.industrydocuments.ucsf.edu/docs/hygl0001

Malloy, M. et alia: 'Exposure to a chloride-deficient formula during infancy: Outcomes at ages 9 and 10 years', *Pediatrics*, 86, 601-610 (1990).

Myers, W. O. et alia. 'Medical versus early surgical therapy in patients with triple-vessel disease and mild angina pectoris: A CASS registry of survival', *Annals of Thoracic Surgery*, 44, 471-486 (1987).

Phillips, C. D., Sloane, P. D., Hawes, C., Koch, G. et alia. 'Effects of residence in Alzheimer disease special care units on functional outcomes', *Journal of the American Medical Association*, 278, 1340-1344 (1997).

Raby, K. E., Goldman, L., Creager, M. A., Cook, E. F. et alia. 'Correlation between preoperative ischemia and major cardiac events after peripheral vascular surgery', *The New England Journal of Medicine*, 321, 1296-1300 (1989).

Reinisch, J., Sanders, S., Mortensen, E. and Rubin, D. B. 'In utero exposure to phenobarbital and intelligence deficits in adult men', *Journal of the American Medical Association*, 274, 1518-1525 (1995).

Soumeri, S. B., McLaughlin, T. J., Spiegelman, D., Hertzmark, E. et alia. 'Adverse outcomes of underuse of beta-blockers in elderly survivors of acute myocardial infarction', *Journal of the American Medical Association*, 277, 115-121 (1997).

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N. et alia. 'Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia', *Medical Care*, 33, AS56-AS66 (1995).

Van Den Eeeden, S. K., Glasser, M., Mathias, S. D., Colwell, H. H. et alia. 'Quality of life, health care utilization, and costs among women undergoing hysterectomy in a managed-care setting', *American Journal of Obstetrical Gynecology*, 178, 91-100 (1998).

Willoughby, A. et alia. 'Population-based study of the development outcome of children exposed to chloride-deficient infant formula', *Pediatrics*, 85, 485-490 (1990).

General Accounting Office. 'Breast conservation versus mastectomy: Patient survival in day-to-day medical practice and randomized studies', Report #GAO-PEMD-95-9, United States General Accounting Office, Washington D.C., 1994.

The SUPPORT Principal Investigators. 'A controlled trial to improve care for seriously ill hospitalized patients', *Journal of the American Medical Association*, 274, 1591-1598 (1995).

CONFIDENTIAL

52299 4744

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Industry Documents Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

Rubin
EXHIBIT NO. 3
7-2-99
J. WILLIAMS

BIOMETRICS 48, 479-495
June 1992

Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders

James M. Robins and Steven D. Mark
Harvard School of Public Health, 665 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A.
and
Whitney K. Newey
Department of Economics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, U.S.A.

Produced by RJRTC
HEY

SUMMARY

In order to estimate the causal effects of one or more exposures or treatments on an outcome of interest, one has to account for the effect of "confounding factors" which both covary with the exposures or treatments and are independent predictors of the outcome. In this paper we present regression methods which, in contrast to standard methods, adjust for the confounding effect of multiple continuous or discrete covariates by modelling the conditional expectation of the exposures or treatments given the confounders. In the special case of a univariate dichotomous exposure or treatment, this conditional expectation is identical to what Rosenbaum and Rubin have called the propensity score. They have also proposed methods to estimate causal effects by modelling the propensity score. Our methods generalize those of Rosenbaum and Rubin in several ways. First, our approach straightforwardly allows for multivariate exposures or treatments, each of which may be continuous, ordinal, or discrete. Second, even in the case of a single dichotomous exposure, our approach does not require subclassification or matching on the propensity score so that the potential for "residual confounding," i.e., bias, due to incomplete matching is avoided. Third, our approach allows a rather general formalization of the idea that it is better to use the "estimated propensity score" than the true propensity score even when the true score is known. The additional power of our approach derives from the fact that we assume the causal effects of the exposures or treatments can be described by the parametric component of a semiparametric regression model. To illustrate our methods, we reanalyze the effect of current cigarette smoking on the level of forced expiratory volume in one second in a cohort of 2,713 adult white males. We compare the results with those obtained using standard methods.

1. Introduction

1.1 *The Problem*

In order to estimate the causal effect of one or more exposures or treatments on an outcome of interest, one has to account for the effect of "confounding factors" which both covary with the exposures or treatments and are independent predictors of the outcome. If few in number, categorical confounding factors are commonly dealt with by stratification. When there are many confounding factors or when some of the factors are continuous, regression methods are used. In this paper we present regression methods which, in contrast to standard methods, adjust for confounding by modelling aspects of the marginal association of the exposures of interest with the confounders rather than by modelling the independent

Key words: Causal inference; Covariance adjustment; Epidemiologic methods; Propensity score; Semiparametric efficiency; Semiparametric regression.

52299 4745

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:31 PM FAX NO. P. 07

association of the confounders with the outcome. Specifically, we will model the conditional expectation of the exposures given the confounders. These methods of estimation will be particularly useful when prior knowledge regarding the association of the confounders with exposure status is more precise than knowledge regarding their association with the outcome.

For concreteness, we shall attempt to estimate the effect of being a current cigarette smoker on the level of forced expiratory volume in one second (FEV1) in a cohort of 2,713 adult white male former and current cigarette smokers from the initial cross-sectional data collected in the Harvard Six Cities Study (Dockery et al., 1988). We shall estimate this effect while adjusting for the presence of the 22 potential confounding factors listed in Table 1 that include past smoking history, past respiratory symptoms, age, height, and coexistent heart disease. In this example the exposure of interest is dichotomous and we assume that there is no interaction between that exposure and the confounders. That is, we assume that the absolute effect of current smoking on FEV1 does not depend on a subject's age, weight, previous smoking history, etc. In this setting the most common approach to estimating the effect of current smoking on FEV1 would be to postulate a linear regression model

$$Y_i = \beta_1 + \beta S_i + \sum_{k=2}^{K} \beta_k X_{k,i} + \epsilon_i, \quad E[\epsilon_i | S_i, X_i] = 0, \tag{1}$$

where Y_i, S_i, $X_i = (X_{2,i}, \ldots, X_{K,i})$ are respectively random variables representing subject i's FEV1 level, current smoking status ($S_i = 1$ if a current smoker and $S_i = 0$ otherwise), and values on a vector X_i of potential confounding factors. Note that the parameter of interest, β, is distinguished from the "nuisance" parameters $(\beta_1, \ldots, \beta_K)$ by the absence of a subscript. For notational simplicity, we shall assume that (Y_i, S_i, X_i) are independent and identically distributed random vectors, although, with minor modifications, our results will hold if the X_i are fixed constants and the (ϵ_i, S_i) are independent across subjects.

Define $\sigma^2(S, X) = \text{var}[\epsilon_i | S, X]$. We write $\sigma^2(S, X) = \sigma^2$ if the errors ϵ_i are homoscedastic. Unless stated otherwise, we shall assume homoscedastic errors, although we do not assume that this fact is known to the data analyst. The ϵ_i are not assumed to be independent of the (S_i, X_i).

Suppose we are unwilling to assume that the independent association of the confounders X_i with the outcome Y_i has a known functional form. In that case, we would generalize model (1) to

$$Y_i = \beta S_i + h(X_i) + \epsilon_i, \quad E[\epsilon_i | S_i, X_i] = 0, \tag{2}$$

where $h(X_i)$ is an unknown real-valued function of the vector X_i. Model (2) has a

Table 1
Twenty-two potential confounders of the effect of current smoking on FEV1

Age	History of emphysema
Age-squared	Past history of asthma
Height	Current asthma
Body mass index	Former cigar smoker
Chronic cough	Current cigar smoker level = hi
Recurrent bouts of coughing	Current cigar smoker level = medium
History of treatment for heart disease	Current cigar smoker level = lo
	Former pipe smoker
Chronic phlegm production	Current pipe smoker level = hi
Chronic wheeze	Current pipe smoker level = medium
Total years of cigarette smoking	Current pipe smoker level = lo
Lifetime pack-years smoked	

52299 4746

2 of 17

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Document hosted at http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:32 PM FAX NO. P. 08

semiparametric regression function with parametric component βS_i and nonparametric component $h(X_i)$. This paper is concerned with the estimation of β from model (2). Robinson (1988) has provided an asymptotically normal and unbiased estimator of β under a large-sample limiting model in which the number of confounding factors remains fixed as the sample size grows. His estimator relies on the fact that, under such a limiting model, the unknown function $h(X_i)$ can be consistently estimated by nonparametric regression techniques. In epidemiologic research, the number of confounding variables can be quite large. In these instances, the more appropriate limiting model would be one in which we allowed the number of confounding factors contained in X_i to increase with the sample size (Huber, 1981).

It is difficult to generalize Robinson's approach based on nonparametric estimation of $h(X_i)$ when the dimension of X_i is large. As a consequence, to obtain consistent estimators of β, we shall consider making additional a priori assumptions beyond those specified by model (2). The standard approach would be to assume that $h(X_i)$ is known a priori except for a finite number of unknown parameters. As an example, the linear regression model (1) assumes that

$$h(X_i) = \beta_1 + \sum_{k=2}^{K} \beta_k X_{k,i}.$$

In contrast to the standard approach, in this paper we shall suppose that prior information concerning the marginal association of S_i with X_i is sharper than that concerning the form of $h(X_i)$. Thus we shall leave $h(X_i)$ completely unspecified and instead specify parametric models for the marginal association of S_i and X_i. Specifically, we shall consider parametric models for $E[S_i|X_i] = p(S_i = 1|X_i)$ such as the logistic regression model

$$p[S = 1|X_i; \alpha] = \frac{\exp(\alpha_1 + \sum_{k=2}^{K} \alpha_k X_{k,i})}{1 + \exp(\alpha_1 + \sum_{k=2}^{K} \alpha_k X_{k,i})}, \tag{3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)$. We shall show that we can obtain asymptotically normal and unbiased estimators of β in model (2) provided our model (3) for $p(S = 1|X_i)$ is correctly specified.

Although correctly specified parametric models for either $h(X_i)$ or $p(S = 1|X_i)$ will provide asymptotically normal and unbiased estimates of β, nonetheless, as discussed in the next paragraph, least squares estimators of β based on models for $h(X_i)$ will always be at least as efficient as any estimator of β based on models for $p(S = 1|X_i)$. This suggests that, for reasons of efficiency, it is always preferable to model $h(X_i)$ rather than $p(S = 1|X_i)$. But if, as we assume in this paper, our prior information concerning $h(X_i)$ is less sharp than that concerning $p(S = 1|X_i)$, we would choose not to model $h(X_i)$ in order to protect against specification bias.

In order to explain why the ordinary least squares estimator of β based on a correctly specified model for $h(X_i)$ is always at least as efficient as any estimator of β based on models for $E[S|X_i]$, we need to review some results from the theory of semiparametric efficiency bounds derived by Chamberlain (1987; and Discussion Paper 1494, Harvard Institute of Economic Research, 1990) and exposited by Newey (1990). For the moment suppose again that, as in equation (1), we were able to correctly specify a parametric model, say, $q(X_i; \theta)$ for $h(X_i)$ depending on a parameter vector θ. In equation (1), $\theta = (\beta_1, \ldots, \beta_K)$. Chamberlain (1987) showed that the estimator of β obtained by fitting the model $Y_i = \beta S_i + q(X_i; \theta) + \epsilon_i$ by unweighted, possibly nonlinear, least squares is the most efficient possible estimator of β that is guaranteed to be asymptotically normal and unbiased under the sole prior restrictions that $E[\epsilon_i|S_i, X_i] = 0$ and $h(X_i) = q(X_i; \theta)$. [If, as in equation (1), $q(X_i; \theta)$ is linear in θ, we fit using ordinary least squares. Otherwise, we fit using nonlinear least

52299 4747

3 of 17

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:34 PM FAX NO. P. 09

squares.] Therefore if, as in model (2), we are unwilling to specify a parametric form for $h(X_i)$ and yet want our estimator of β to be asymptotically normal and unbiased whatever be $h(X_i)$, the asymptotic variance of any such estimator clearly cannot be less than the supremum of the asymptotic variances of the least squares estimators of β taken over the set of all possible parametric models for $h(X_i)$. This supremum is called the semiparametric efficiency bound for an estimator of β under model (2) (Bickel et al., 1992) and was shown by Chamberlain (discussion paper cited previously) to equal $n^{-1}\sigma^2/E[\text{var}(S|X)]$, where n is the sample size.

Thus, if we are able to correctly specify a parametric model for $h(X_i)$, the least squares estimator of β always has variance no greater than the efficiency bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. In contrast, if under model (2), we are unable to specify a parametric model for $h(X_i)$, but instead correctly specify a model for $E[S|X_i]$, no estimator that is asymptotically unbiased for β for all $h(X_i)$ can have variance less than the bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. This is a consequence of the fact that $\{(S_i, X_i), i \in (1, \ldots, n)\}$ is ancillary for β under model (2) (Cox and Hinkley, 1974) and, as discussed by Newey (1990), knowledge concerning the marginal distribution of an ancillary statistic does not affect the semiparametric efficiency bound for the estimation of β.

It needs to be stressed that, even when we can obtain a consistent estimator of β in model (2), it does not follow that the parameter β can be interpreted as the causal effect of current cigarette smoking on FEV1. We now describe conditions under which β does have a causal interpretation.

1.2 A Causal Model

Following Rubin (1978), let $Y_{S=1,i}$ be subject i's FEV1 had subject i been a current smoker. If subject i is a current smoker in the actual study, then $Y_{S=1,i}$ equals his observed FEV1 Y_i. If subject i is not a current smoker, $Y_{S=1,i}$ is missing. Similarly, $Y_{S=0,i}$ is subject i's FEV1 if subject i were, possibly contrary to fact, a current nonsmoker. Rubin defined the average causal effect of current smoking among subjects with observed covariates level X_i to be $E[Y_{S=1}|X_i] - E[Y_{S=0}|X_i]$. Now, under our model (2), we know that $E[Y|X_i, S = 1] - E[Y|X_i, S = 0] = \beta$ since $E[Y|X_i, S = 1] = \beta + h(X_i)$ and $E[Y|X_i, S = 0] = h(X_i)$.

Thus a sufficient condition for β to equal the average causal effect of current smoking at each level X_i is that, for each X_i,

$$E[Y_{S=s}|X_i] = E[Y|X_i, S = s], \quad s \in \{0, 1\}. \tag{4a}$$

Under Rubin's causal model, equation (4a) is equivalent to

$$E[Y_{S=s}|X_i] = E[Y_{S=s}|X_i, S = s]. \tag{4b}$$

We shall assume that equation (4b) holds and thus β has a causal interpretation when X_i is the vector of 22 potential confounding variables described above. The assumption that equation (4b) holds is nonidentifiable in the sense that it is compatible with any joint distribution for the observable random variables (S_i, X_i, Y_i). When equation (4b) holds, we shall call model (2) a *semiparametric causal regression model*. Equation (4b) says that, conditional on the joint level of the 22 potential independent risk factors X_i, the mean of $Y_{S=s}$ among subjects who actually receive treatment $S = 1$ equals that among subjects who actually receive treatment $S = 0$. We do not assume that equation (4b) holds when X_i is a proper subset of the 22 potential confounding variables.

The mathematical results in this paper are concerned only with the estimation of β in model (2) and do not depend on whether equation (4b) holds. Of course, in general, we are interested in the estimation of β only when we believe it has a causal interpretation.

Produced by R.J.R.T.C.

52299 4748

4 of 17

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:35 PM FAX NO. P. 10

1.3 *Relationship to the Propensity Score*

Rosenbaum and Rubin (1983, 1984, 1985) and Rosenbaum (1984, 1987, 1988) have also considered estimating the causal effect of a dichotomous treatment such as S_i on an outcome Y_i by modelling $p(S = 1|X_i)$ when equation (4) holds. These authors call $p[S = 1|X_i]$ the propensity score. In contrast to their approach, our approach straightforwardly allows the treatment or exposure S_i to be continuous or ordinal rather than simply dichotomous. Furthermore, as discussed in the Appendix, our approach allows S_i to be multivariate so that we can, say, estimate the independent effects of current cigarette smoking and past cigarette smoking. In addition, our "regression" approach does not require subclassification or matching on the propensity score $p[S = 1|X_i]$ even when X_i has continuous components so that the potential for "residual" confounding, i.e., bias, due to the fact that one has not precisely matched on $p[S = 1|X_i]$ is avoided. The additional power of our approach derives from the fact that we assume the causal effect of exposure can be described by the parametric component of a semiparametric causal regression model such as model (2).

Rosenbaum (1984, 1988) also considered specifying causal models to avoid the need to match or subclassify on the propensity score. In general, Rosenbaum is concerned with small-sample (exact) rather than large-sample (asymptotic) inference. As a consequence, his causal models tend to be even more restrictive than model (2). Specifically, he assumes a constant treatment effect model—that is, $Y_{S=1,i} = \beta + Y_{S=0,i}$ for all subjects i—although his results would still hold under the weaker assumption that the distributions of $Y_{S=1,i}$ and $Y_{S=0,i}$ differed by a "shift" parameter β. Furthermore, as he points out, his "exact" methods do not allow one to adjust for the confounding effects of continuous covariates.

Finally, as discussed in Section 2, our approach allows a rather general formalization of the idea that it is better to use the "estimated" propensity score than the "true" propensity score even when the true score is known (Rosenbaum, 1987).

2. Estimators Based on Models for the Conditional Expectation of Exposure Given Confounders

2.1 *An Infeasible Estimator*

In this section, we consider estimators of β under model (2) when we can specify accurate models for $E(S|X_i)$. Note that when S_i is dichotomous, models for $E(S|X_i)$ are models for $p(S = 1|X_i)$. Initially, for pedagogic purposes, we shall assume that we know $E(S|X_i)$ exactly. That is, we assume exact prior knowledge of the expected value of S for every combination of the confounders X_i. Subsequently we make the more tenable assumption that we know $E(S|X_i)$ up to a finite vector of unknown parameters. We allow $\sigma^2(S, X)$ to depend on (S, X). Henceforth, we adopt the following notational convention: β will refer to the true but unknown value of the coefficient of S_i in model (2); β^T will refer to any hypothesized, possibly incorrect, value for β.

The estimator we shall consider, which we call the E-estimator,

$$\hat{\beta}_E = \frac{\sum_{i=1}^{n} Y_i[S_i - E(S|X_i)]}{\sum_{i=1}^{n} S_i[S_i - E(S|X_i)]}, \tag{5}$$

is based on a suggestion by Newey (1990).

It is shown in Theorem A.1 in the Appendix that $\hat{\beta}_E$ has a limiting normal distribution with mean β.

The consistency of $\hat{\beta}_E$ is based on the fact that model (2) implies that

$$E[z_i|X_i, S_i] = E[z_i|X_i]. \tag{6}$$

52299 4749

5 of 17

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf Stanford Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:36 PM FAX NO. P. 11

where $z_i = Y_i - S_i\beta$. In the proof of Theorem A.1 in the Appendix, it is shown that equation (6) implies the identity

$$E[U(\beta)] = 0, \tag{7}$$

where, for any $\beta^\dagger$, $U(\beta^\dagger) = \sum_{i=1}^{n} (Y_i - S_i\beta^\dagger)(S_i - E(S|X_i))$. The E-estimator $\hat{\beta}_E$ is the solution $\beta^\dagger$ to the unbiased estimating equation $U(\beta^\dagger) = 0$.

2.2 *A Feasible Estimator*

Of course, the estimator $\hat{\beta}_E$ is not feasible since, in practice, $E(S|X_i)$ is unknown. We can overcome this difficulty if we assume a priori that the logistic regression model equation (3) holds. We then estimate $E(S|X_i)$ by logistic regression and subsequently estimate β by

$$\tilde{\beta}_E = \frac{\sum_{i=1}^{n} Y_i[S_i - \hat{E}(S|X_i)]}{\sum_{i=1}^{n} S_i[S_i - \hat{E}(S|X_i)]}, \tag{8}$$

where $\hat{E}(S|X_i)$ is the fitted value $\hat{p}_i = p[S = 1|X_i; \hat{\alpha}]$ of $p[S = 1|X_i]$, and $\hat{\alpha}$ is the maximum likelihood estimator of α from the logistic regression. Note that we use the symbol $\tilde{\beta}_E$ rather than $\hat{\beta}_E$ to represent the feasible estimator of equation (8).

As shown in Theorem A.1 in the Appendix, it follows from Pierce (1982) and Newey (1990) that when the logistic model of equation (3) is true, $\tilde{\beta}_E$ is asymptotically normal and unbiased and its asymptotic covariance matrix can be consistently estimated by

$$\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E) = \text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E) - \tilde{Q}[\text{var}_{\text{est}}^{\wedge}(\hat{\alpha})]\tilde{Q}^T \tag{9}$$

where

$$\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E) = \frac{\sum_{i=1}^{n} \hat{z}_i^2(S_i - \hat{p}_i)^2}{[\sum_{i=1}^{n} S_i(S_i - \hat{p}_i)]^2}, \tag{10}$$

$\hat{z}_i = Y_i - \hat{\beta}_E S_i$, $\tilde{Q}^T$ is the K-vector with jth component

$$\tilde{Q}_j = \frac{-\sum_{i=1}^{n} \hat{z}_i \hat{p}_i(1 - \hat{p}_i)X_{i,j}}{\sum_{i=1}^{n} S_i(S_i - \hat{p}_i)}$$

(where we define $X_{i,1} = 1$ when $j = 1$), and $\text{var}_{\text{est}}(\hat{\alpha})$ is the estimated covariance matrix (i.e., the inverse of the observed information matrix) from the fit of the logistic model equation (3). The observed information matrix has (j, k) entry $-\sum_{i=1}^{n} \hat{p}_i(1 - \hat{p}_i)X_{i,j}X_{i,k}$. The estimator $\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E)$ is not guaranteed to be positive-definite. A positive-definite consistent variance estimator is obtained by replacing $\hat{p}_i(1 - \hat{p}_i)$ by $(S_i - \hat{p}_i)^2$ both in the numerator of $\tilde{Q}_j$ and in the observed information matrix.

Even though $\hat{\beta}_E$ is infeasible when $p[S = 1|X_i]$ is unknown, $\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E)$ is still a feasible consistent estimator of its asymptotic variance. Therefore it follows from equation (9) that one generates a more precise estimate of β by *estimating* the propensity score $E(S|X_i)$ than by using the true population value of the propensity score even were the latter known. That is, $\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E)$ is always less than or equal to $\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E)$. As discussed in the Appendix, this result depends on the fact that $\hat{\alpha}$ is an efficient estimator of α. The preference for $\tilde{\beta}_E$ compared to $\hat{\beta}_E$ when the parameter α of model (3) is known can also be viewed in terms of conditional bias. Specifically, it can be shown that, conditional on the ancillary statistic $[\text{var}_{\text{est}}(\tilde{\alpha})]^{-1/2}(\hat{\alpha} - \alpha)$, $\hat{\beta}_E$ becomes asymptotically biased while $\tilde{\beta}_E$ remains asymptotically unbiased (Robins and Morgenstern, 1987; Rosenbaum, 1987; Efron and Hinkley, 1978).

In Table 2 we present four different estimates $\hat{\beta}_E$ of β based on specifying four different

Produced by IMPRIMIS

52299 4750

6 of 17

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-89 THU 08:37 PM FAX NO. P. 12

Table 2
Estimates $\tilde{\beta}_E$ under four different specifications for $p[S = 1 \mid X_t]$

Analysis	Covariates $X_{k,t}$ included in logistic model of equation (3) for $p[S = 1 \mid X_t]$	$\tilde{\beta}_E$	$\widehat{\text{var}}_{cm}(\tilde{\beta}_E) \times 10^{-4}$	$\widehat{\text{var}}_{\infty}(\tilde{\beta}_E) \times 10^{-4}$
(1)	Constant term only	−.0580	9.49	159.0
(2)	Constant, chronic cough (Yes, No)	.0429	9.36	167.0
(3)	Constant, pack-years of smoking	.0520	7.45	157.9
(4)	Constant, 22 covariates in Table 1	−.1133	8.82	332.6

logistic regression models for $p[S = 1 \mid X_t]$. In the first analysis in Table 2, we assume no confounding. That is, we fit only a constant term α_1 in equation (3). In the second analysis X_t in equation (3) is the single binary covariate—history of chronic cough. In the third analysis X_t is the single continuous covariate—lifetime number of pack-years. In the fourth analysis X_t in equation (3) is the 22-vector of potential confounders. The striking efficiency advantage attributable to estimating the propensity score $p[S = 1 \mid X_t]$ can be obtained by comparing $\widehat{\text{var}}_{cm}(\tilde{\beta}_E)$ to $\widehat{\text{var}}_{\infty}(\tilde{\beta}_E)$ in Table 2.

Under the assumptions that (a) the coefficient β in equation (2) has a causal interpretation [i.e., equation (4b) holds] when X_t is the 22-vector of confounders and (b) the model for $p[S = 1 \mid X_t]$ used in analysis (4) is true, analysis (4) provides a consistent estimator of this causal β. Therefore, we estimate that current smoking causes a decrease of .1133 liter in FEV1. A 95% confidence interval for β is $-.113 \pm (1.96)(.00088)^{1/2} = (-.170, -.056)$.

Under assumptions (a) and (b), we now provide sufficient conditions for the simpler analyses (1)-(3) also to provide consistent estimators of the "causal" β associated with model (2) with X_t the 22-vector of covariates.

We shall restrict attention to analysis (3) since the conditions for analyses (1) and (2) are similar. Let X_{k^*} be the covariate "lifetime number of pack-years" used in analysis (3). $\tilde{\beta}_E$ from analysis (3) will be consistent for the causal β if either of the following is true:

Sufficient condition (1): With X_t the 22-vector of covariates, $\alpha_k = 0$ for the 21 covariates X_k in the logistic model (3) other than X_{k^*} (i.e., lifetime pack-years is the only predictor of current smoking among the 22 potential confounding factors).

Sufficient condition (2): The unknown function $h(X_t) = h(X_{2,t}, \ldots, X_{K,t})$, $K = 22$, is actually only a function of $X_{k^*,t}$ (i.e., lifetime pack-years is the only independent risk factor among the 22 potential confounding factors) and $p[S = 1 \mid X_{k^*,t}]$ follows a linear logistic model.

In general it would be unlikely that an investigator would be willing to assume that either of the above sufficient conditions held, and thus would tend to rely on analysis (4).

Suppose equation (4b) holds and consider the test of the null hypothesis $\beta = 0$ that rejects if $\tilde{\beta}_E \pm 1.96[\widehat{\text{var}}_{cm}(\tilde{\beta}_E)]^{1/2}$ fails to include 0. Then except for the assumption that the model (3) for $p[S = 1 \mid X_t]$ is correctly specified, this test is an "otherwise asymptotically distribution-free" .05 α-level test of the sharp null hypothesis of no causal effect of exposure, i.e., of the hypothesis $Y_{S=1,t} = Y_{S=0,t} = Y_t$ for all subjects t.

Rosenbaum (1984, §4.2) proposes a test of this null hypothesis that will be "otherwise asymptotically distribution-free" under the condition that $(Y_{S=1,t}, Y_{S=0,t})$ and S_t are conditionally independent given X_t.

52299 4751

7 of 17

http://legacy.library.ucsf.edu/tid/eoq07a00/pdf Powered by www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:38 PM FAX NO. P. 13

3. Relationship of E-Estimators to Ordinary Least Squares

The ordinary least squares (OLS) estimator of β in equation (1) can be written

$$\tilde{\beta}_{OLS} = \frac{\sum Y_i(S_i - \hat{P}(S|X_i))}{\sum [S_i - \hat{P}(S|X_i)]^2}, \tag{11}$$

where summation signs without indexes will refer to sums over individuals and where $\hat{P}(S|X_i)$ is the fitted value from the OLS regression of S on X_i and the constant one. Now the right-hand side of equation (11) can be written as

$$\tilde{\beta}_{OLS} = \frac{\sum Y_i(S_i - \hat{P}(S|X_i))}{\sum S_i(S_i - \hat{P}(S|X_i))}$$

using the fact that, for OLS, the empirical correlation of the fitted values and the residuals is zero. Now suppose we had modelled $E(S|X_i) = p[S = 1|X_i]$ by the linear probability model $p[S = 1|X_i, \alpha] = \alpha_1 + \sum_{k=2}^{K} \alpha_k X_{k,i}$ rather than by a logistic model, and we fit the linear probability model by least squares. Then $\hat{E}(S|X_i) = \hat{P}(S|X_i)$. Therefore, from its definition $\tilde{\beta}_E = \tilde{\beta}_{OLS}$. In the previous section we showed that $\tilde{\beta}_E$ is consistent if our model for $E(S|X_i)$ is true. It follows, as pointed out by Newey (1990), that if, in truth, $E(S|X_i) = \alpha_1 + \sum_{k=2}^{K} \alpha_k X_{k,i}$ [i.e., $E(S|X_i)$ is linear in X_i], then $\tilde{\beta}_{OLS}$ is consistent for β even if $h(X_i)$ is nonlinear and thus equation (1) is false. Nonetheless if $h(X_i)$ is nonlinear, the estimate of the variance of $\tilde{\beta}_{OLS}$ provided by standard software packages is inconsistent, and equation (9) must be used. If $E(S|X_i)$ is not linear in X_i, the ordinary least squares estimate of β would, in general, be inconsistent if the unknown function $h(X_i)$ is, in truth, nonlinear in X_i.

Table 3 shows $\tilde{\beta}_{OLS}$ from the fit of equation (1) for the four choices of X_i as in Table 2. Note that $\tilde{\beta}_{OLS} = \tilde{\beta}_E$ in analysis (2). This reflects the fact that when X_i is a single dichotomous covariate, $E(S|X_i)$ is simultaneously linear and linear logistic, and $\hat{E}(S|X_i) = \hat{P}(S|X_i)$. For similar reasons $\tilde{\beta}_{OLS} = \tilde{\beta}_E$ in analysis (1). $\tilde{\beta}_E$ from analysis (3) using the continuous variable "pack-years" is not identical to the OLS estimate since $\hat{E}(S|X_i) \neq \hat{P}(S|X_i)$. The fact that $\tilde{\beta}_E$ and $\tilde{\beta}_{OLS}$ are close can be explained by the near linearity of $\hat{E}(S|X_i)$ in our data, which can be checked by plotting $\hat{E}(S|X_i)$ versus X_i.

We now discuss a modification of the estimator $\tilde{\beta}_E$ that has an even closer connection to OLS than does $\tilde{\beta}_E$. Define

$$\tilde{\beta}_{Em} = \frac{\sum Y_i(S_i - \hat{E}(S|X_i))}{\sum [S_i - \hat{E}(S|X_i)]^2}. \tag{12}$$

When $E(S|X_i)$ is nonlinear (e.g., logistic), $\tilde{\beta}_{Em}$ will not in general equal $\tilde{\beta}_E$. Nonetheless,

Table 3
Estimates $\tilde{\beta}_{OLS}$ under four different specifications for covariates included in $\beta_1 + \sum \beta_k X_{k,i}$ in model equation (1)

Analysis	Covariates $X_{k,i}$ included in equation (1)	$\tilde{\beta}_{OLS}$	$\text{var}_m(\tilde{\beta}_{OLS})^a \times 10^{-4}$
(1)	Constant term only	−.0580	9.50
(2)	Constant, chronic cough (Yes, No)	.0429	9.39
(3)	Constant, pack-years of smoking	.0492	7.64
(4)	Constant, 22 covariates in Table 1	−.1199	8.68

[a] Using White's (1980) heteroscedastic consistent variance estimator.

52299 4752

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:39 PM FAX NO. P. 14

$\tilde{\beta}_{Em}$ and $\tilde{\beta}_E$ have the same asymptotic distribution. One obtains $\tilde{\beta}_{Em}$ by regressing Y_i versus "the residuals" $S_i - \hat{E}(S|X_i)$ using OLS regression with no intercept.

4. Two-Stage E-Estimators

Throughout this section we assume that the logistic model equation (3) is correctly specified with X_i the vector of 22 covariates. Then $\tilde{\beta}_E$ is a consistent estimator of β in equation (2) without making any assumptions about the form of $h(X_i)$. Suppose now that we have an a priori guess as to the shape of $h(X_i)$. For concreteness, suppose we believed that $h(X_i)$ was linear or at least nearly linear in X_i, i.e., $h(X_i) = \beta_1 + \sum_{k=2}^{\kappa} \beta_k X_{k,i}$. We can now consider how to develop an estimator, say $\tilde{\beta}^*$, that may be much more efficient than $\tilde{\beta}_E$ if our guess concerning the shape of $h(X_i)$ is correct or nearly correct, and will remain consistent, asymptotically normal no matter how wrong our guess may be. To construct $\tilde{\beta}^*$, we proceed in two steps. First we compute $\hat{E}(S|X_i)$ and $\tilde{\beta}_E$ as before. We then regress that $\tilde{z}_i = Y_i - \tilde{\beta}_E S_i$ on X_i. We then define $\tilde{\beta}^*$ to be the solution $\beta^\dagger$ to the estimating equation

$$0 = \Psi^*(\beta^\dagger) \equiv \sum \left(Y_i - \beta^\dagger S_i - \tilde{\beta}_1 - \sum_{k=2}^{\kappa} \tilde{\beta}_k X_{k,i} \right)(S_i - \hat{E}(S|X_i)),$$

where $(\tilde{\beta}_1, \ldots, \tilde{\beta}_\kappa)$ are the OLS estimates from the regression of $\tilde{z}_i$ on X_i. Therefore,

$$\tilde{\beta}^* = \frac{\sum (Y_i - \tilde{\beta}_1 - \sum_{k=2}^{\kappa} \tilde{\beta}_k X_{k,i})[S_i - \hat{E}(S|X_i)]}{\sum S_i(S_i - \hat{E}(S|X_i))}.$$

In Theorems A.1 and A.3 in the Appendix we show that $\tilde{\beta}^*$ is asymptotically normal and unbiased even if the proposed linear model for $h(X_i)$ is incorrect.

A consistent estimator of $\text{var}^\wedge(\tilde{\beta}^*)$ is

$$\frac{\sum \tilde{u}_i^2 (S_i - \hat{p}_i)^2}{[\sum S_i(S_i - \hat{p}_i)]^2} - (\hat{Q}^*)\text{var}_{\hat{\alpha}}(\hat{\alpha})(\hat{Q}^*)^T, \tag{13}$$

where $\tilde{u}_i = Y_i - \tilde{\beta}^* S_i - \sum_{k=2}^{\kappa} \tilde{\beta}_k X_{k,i} - \tilde{\beta}_1$ and $(\hat{Q}^*)^T$ has components

$$\hat{Q}_j^* = \frac{-\sum \tilde{u}_i \hat{p}_i(1 - \hat{p}_i) X_{j,i}}{\sum S_i(S_i - \hat{p}_i)}.$$

In our example $\tilde{\beta}^*$ is $-.117$ with $\text{var}\hat{\wedge}_n(\tilde{\beta}^*) = 8.79 \times 10^{-4}$ when X_i is the 22-vector of covariates.

In the final paragraph of the Appendix we show that, if the linear model postulated for $h(X_i)$ were correct, then (1) $\hat{Q}^*$ converges to zero in probability so the correction term could be ignored; and (2) if $\sigma^2(S, X) = \sigma^2$, then $\text{var}^\wedge(\tilde{\beta}^*) = n^{-1}\sigma^2/E[\text{var}(S|X)]$.

When $h(X_i)$ is linear, $\tilde{\beta}_{OLS}$ will be consistent asymptotically normal and $\text{var}^\wedge(\tilde{\beta}_{OLS})$ will be less than or equal to $\text{var}^\wedge(\tilde{\beta}^*)$, with equality when $E(S|X_i)$ is linear in X_i. Of course, if neither $h(X_i)$ nor $E(S|X_i)$ is linear, $\tilde{\beta}^*$ but not $\tilde{\beta}_{OLS}$ remains consistent [provided the nonlinear model for $E(S|X_i)$ is correctly specified]. When $\sigma^2(S, X) = \sigma^2$ and $h(X_i)$ is, in truth, linear, $\tilde{\beta}^*$ has the smallest asymptotic variance among all estimators that remain asymptotically unbiased even were $h(X_i)$ nonlinear (Chamberlain, discussion paper cited previously). That is, it attains the semiparametric efficiency bound for model (2).

5. Discussion

Suppose again that β in model (2) is causal [i.e., equation (4b) holds] when X_i is the 22-vector of covariates. Then the validity of our E-estimators of the causal effect of current smoking on FEV1 requires that the semiparametric regression model (2) and logistic

Produced by RJR TIC
UNMIN
52299 4753
9 of 17
http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Univ. http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:40 PM FAX NO. P. 15

regression (3) be correctly specified. Specification of (3) can be checked using the techniques described by Landwehr, Pregibon, and Shoemaker (1984). The no-interaction assumption of model (2) can be checked by nesting (2) in the more general semiparametric regression model of the Appendix that includes interactions between current smoking S_i and the covariates in X_i, and then testing whether the interaction coefficients are nonzero.

We note that, rather than simply modelling $p[S = 1|X_i]$ by the linear no-interaction logistic model equation (3), we could continue to add to equation (3) additional terms such as powers of the $X_{k,i}$ (e.g., $X_{k,i}^2$, $X_{k,i}^3$) for continuous covariates and all orders of interaction between the various covariates and their powers (e.g., $X_{3,i}^3 \cdot X_{4,i} \cdot X_{5,i}^3$). This will greatly increase the number of free coefficients in our model for $p[S = 1|X_i]$. As we add these additional terms, we derive two benefits. First, we decrease any asymptotic bias in $\hat{\beta}_E$ (or $\hat{\beta}^*$) due to possible misspecification of the linear no-interaction model for $p[S = 1|X_i]$. Second, when the linear no-interaction logistic model is *correctly specified* and thus the additional terms are not necessary to make $\hat{\beta}_E$ unbiased, the asymptotic variance of $\hat{\beta}_E$ (or $\hat{\beta}^*$) is nonincreasing and will usually decrease as the number of free parameters in the model for $p[S = 1|X_i]$ increases (see Pierce (1982) and Corollary A.1 of the Appendix). Thus, rather than having the usual tradeoff between efficiency and bias, we find that increasing the number of free parameters can lead to improvements in both bias and efficiency. This apparent "free lunch" must be tempered by two facts. First, no matter how many terms we add, $\widehat{\text{var}}(\hat{\beta}_E)$ and $\widehat{\text{var}}(\hat{\beta}^*)$ will always exceed $n^{-1}\sigma^2/E[\text{var}(S|X)]$ (with homoscedastic errors) (Chamberlain, 1987). Second, the results we have derived require that the estimates of the free parameters in the model for $p[S = 1|X_i]$ are $n^{1/2}$-consistent. [Newey (1990) suggests that $n^{1/4}$-consistency is sufficient.] This limits the number of free parameters we may have in our model for $p[S = 1|X_i]$ as a function of sample size. For example, we could not allow the number of free parameters to equal the total sample size. Cross-validation techniques for model selection should be useful in choosing a proper ratio of sample size to parameters. Moderate and small-sample simulation studies are needed as a guide to practice.

We note that when the linear no-interaction logistic model (3) is *misspecified*, the asymptotic variance of the (now potentially biased) estimator $\hat{\beta}_E$ based on a misspecified model for $p[S = 1|X_i]$ can be less than the asymptotic variance of the estimator $\hat{\beta}_E$ based on a more richly parameterized, correctly specified model in which the misspecified model is nested. This phenomenon is evident in a comparison of analyses (3) and (4) in Table 2. The estimated variance of $\hat{\beta}_E$ in analysis (3) is less than that in analysis (4), because covariates other than "pack-years of smoking" are also important predictors of current smoking.

The results described in the preceding three paragraphs help to clarify both when E-estimation will and will not be preferable to standard covariance adjustment by least squares. Consider first the case in which the sample size is quite large and the dimension of X_i is small, so that richly parameterized models for either $h(X_i)$ or $p[S = 1|X_i]$ can be used. Then, as discussed above and in technical detail by Newey (1990), as one adds power and interaction terms to the model (3) for $p[S = 1|X_i]$, any bias in $\hat{\beta}^*$ and $\hat{\beta}_E$ would tend to zero and the asymptotic variance of $\hat{\beta}^*$, and even $\hat{\beta}_E$, will approach the semiparametric efficiency bound of $n^{-1}\sigma^2/E[\text{var}|S(X)]$. Similarly, in this setting, if we expanded the linear regression model (1) by adding additional terms such as powers of $X_{k,i}$ and interactions between the $X_{k,i}$ and their powers, the bias of $\hat{\beta}_{OLS}$ from the least squares fit of (1) would tend to zero, and the variance of $\hat{\beta}_{OLS}$ would approach the efficiency bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. Thus, in this setting, the use of highly parameterized models for $h(X_i)$ fit by least squares or the use of highly parameterized models for $p[S = 1|X_i]$ fit by E-estimation leads to estimators of β with similar properties.

produced by RJRTC

52299 4754

10 of 17

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Generated from www.industrydocuments.ucsf.edu/docs/hygl0001

Now, consider the case in which the dimension of X_i is large and/or the sample size is moderate. One is then restricted to choosing parsimonious parametric models for $h(X_i)$ and/or $p[S = 1 | X_i]$. Further, since the ratio of the sample size to the dimension of X_i is small, the power to discriminate between correct and incorrectly specified models for $h(X_i)$ and/or $p[S = 1 | X_i]$ will be poor. If, as is often the case in an etiologic study, our primary interest is in obtaining valid inferences concerning β (e.g., confidence intervals that cover at their nominal rate), it is essential to try to obtain asymptotically unbiased estimators of β. Since, in general, unbiased estimation of β requires that the model used in the analysis be correct, we would prefer E-estimation over least squares estimation if we believed that our ability to specify nearly correct parsimonious models for $p[S = 1 | X_i]$ exceeded our ability to specify such models for $h(X_i)$. This would be the case when the investigator thinks, based on substantive considerations, that his or her knowledge of the shape of the regression surface $p[S = 1 | X_i]$ is sharper than knowledge of the shape of the function $h(X_i)$. In the special case, represented by our example in Section 3, in which the fitted regression surface $\hat{p}[S = 1 | X_i]$ is nearly linear in the X_i, E-estimation and standard covariance adjustment by least squares will provide similar estimates irrespective of whether $h(X_i)$ is or is not linear.

We next consider whether it might be possible to develop robust E-estimators. Even if the linear model (1) were true, the efficiency of $\hat{\beta}_{OLS}$ would be poor if the errors ϵ_i have heavy-tailed distributions (Huber, 1981). If we are willing to assume that, in addition to (1), the errors were independent of the (S_i, X_i), efficient robust estimation based on M, L, or R estimators is possible (Huber, 1981). If ϵ_i is independent of (S_i, X_i) but model (1) were not true, robust E-estimation of model (2) could be based on solving an unbiased estimating equation of the form $\sum_i m(Y_i - \beta' S_i, X_i)(S_i - E[S | X_i]) = 0$, where the function $m(Y_i - \beta' S_i, X_i)$ would be chosen to downweight observations for which $Y_i - \beta S_i$ differs greatly from its expected value given X_i. (Such observations will be associated with large values of the residuals.) How to choose the function $m(Y_i - \beta' S_i, X_i)$ in this setting is outside the scope of this paper.

If $\sigma^2(S, X)$ depends on X alone or on S and X, it is possible to develop "weighted" E-estimators that will be more efficient than the E-estimators $\hat{\beta}_E$ or $\tilde{\beta}^*$ (Chamberlain, 1987).

Suppose next that the outcome of interest is a dichotomous disease variable. Then Y_i will be a Bernoulli random variable. In that case, one might no longer wish to specify the semiparametric model (2), i.e.,

$$E[Y_i | X_i, S_i] = h(X_i) + \beta S_i,$$

since the model does not naturally obey the restriction that probabilities must lie in the interval [0, 1]. Therefore one might specify a semiparametric logistic model

$$E[Y_i | X_i, S_i] = \frac{\exp[h(X_i) + \beta S_i]}{1 + \exp[h(X_i) + \beta S_i]}. \tag{14}$$

Unfortunately, the approach developed in this paper will not allow us to consistently estimate the β of equation (14) even though Bickel et al. (1992) and Chamberlain (discussion paper cited previously) show that, in principle, there should exist an $n^{1/2}$-consistent estimator of β based on data (X_i, S_i, Y_i) [at least when the dimension of X_i is fixed as the sample size increases]. Our approach fails because it is fundamentally based on the fact that, for model (2), β is identified from the "pseudo-data" (S_i, V_i, Y_i), where $V_i = E(S | X_i)$. We call V_i "pseudo-data." For example, if V_i were known, our estimator $\hat{\beta}_E$ does not require data on X_i. It can be shown that β in equation (14) is not identified from pseudo-data (S_i, V_i, Y_i) due to the "noncollapsibility" of the logistic parameter β when we collapse from the "raw

produced by RJRTC

52299 4755

http://legacy.library.ucsf.edu/tid/coq07a00/pdf
Source: https://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:45 PM FAX NO. P. 11

data" X_i to V_i. Indeed, Gail, Wieand, and Piantadosi (1984) essentially prove this nonidentifiability result in the special case for which $V_i = \frac{1}{2}$ for all subjects. In fact, suppose X_i were dichotomous and thus e^{β} was the *common* exposure (S)–disease (Y) odds ratio in the two 2×2 tables indexed by the levels of X. In this special case, the nonidentifiability of β when V_i is a fixed constant for all subjects i is simply a restatement of the following well-known fact. Even when S and X are (marginally) independent, the common odds ratio e^{β} is not identified from data (S_i, Y_i) since the marginal exposure–disease odds ratio (ignoring X) may differ from e^{β} and the magnitude of the difference depends on the distribution of X (Gail et al., 1984). However, in contrast to our nonidentifiability results for the β of model (14), if equation (4b) holds, the average causal effect of S on disease Y, i.e., $E[Y_{S=1}] - E[Y_{S=0}]$, is identified from (S_i, V_i, Y_i) (Rosenbaum and Rubin, 1983).

Suppose next that Y_i has a Poisson or overdispersed Poisson distribution. We might then wish to specify semiparametric log-linear models, e.g.,

$$E[Y_i | X_i, S_i] = \exp[h(X_i) + \beta S_i]. \tag{15}$$

For log-linear models, a simple modification of our approach can be used to consistently estimate β from pseudo-data (S_i, V_i, Y_i). Specifically, since, under model (15), $E[U(\beta)] = 0$, where

$$U(\beta^{\dagger}) \equiv \sum_{i=1}^{n} Y_i e^{-\beta^{\dagger} S_i} (S_i - E(S | X_i)), \tag{16}$$

the solution $\hat{\beta}_E$ to $U(\beta^{\dagger}) = 0$ will be consistent, asymptotically normal. A feasible consistent estimator $\hat{\beta}_E$ can be obtained from data (S_i, X_i, Y_i) by specifying a (correct) model for $E[S | X_i]$.

The methods of E-estimation can be extended to estimate the causal effect of a time-varying treatment. Specifically, Robins (1989a, 1992a, 1992b, 1992c, 1992d) and Robins et al. (1992) use an extension of E-estimation, which they call G-estimation, to estimate, from observational data, the causal effect of a time-varying treatment both on a survival time outcome and on the evolution of the mean of a continuous outcome variable measured repeatedly over time in the presence of time-dependent confounding factors. Robins (1989a, 1992b, 1992d) uses G-estimation to correct for noncompliance in randomized trials studying the effect of a time-varying treatment both on survival time outcomes and on the evolution of the mean of a continuous outcome variable when noncompliance depends on time-dependent prognostic factors. G-estimation is of particular importance in estimating the causal effect of a time-varying treatment in the presence of time-varying prognostic factors because standard covariance adjustment based on time-dependent Cox proportional hazard models for survival time outcomes or generalized estimating equations (Liang and Zeger, 1986) for repeated measures outcomes cannot consistently estimate the treatment effect (Robins, 1986, 1989a, 1989b, 1992a, 1992b, 1992c).

ACKNOWLEDGEMENTS

This work was supported in part by National Institutes of Health Grants 2 P30 ES00002, R01-ES03405, K04-ES00180, ES01108, and CA09001. We would like to thank Doug Dockery, Frank Speizer, Benjamin Ferris, and other contributors to the Harvard Six Cities Study for their generous sharing of time and data.

RÉSUMÉ

Pour estimer l'influence d'un ou plusieurs facteurs sur une variable d'intérêt, il faut prendre en compte les effets des covariables qui d'une part varient avec les dits facteurs, et d'autre part aident à

52299 4756

12 of 17

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf

Source: http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:44 PM FAX NO. P. 18

prédire la variable d'intérêt, indépendamment de ces facteurs. Dans cet article, nous présentons des méthodes de régression qui, à la différence des méthodes usuelles, ajustent l'effet confondant de plusieurs covariables (continues ou discrètes) par modélisation de l'espérance conditionnelle des différents facteurs en fonction des covariables. Dans le cas particulier d'un seul facteur à deux niveaux, cette espérance conditionnelle est identique à ce que Rosenbaum et Rubin ont appelé le score de propension. Ces auteurs, d'ailleurs, ont aussi proposé des méthodes d'estimation passant par la modélisation de ce score de propension. Nos méthodes généralisent celles de Rosenbaum et Rubin de plusieurs manières. Tout d'abord, notre approche s'étend d'emblée à tous les cas de figure possibles pour les facteurs, chacun d'entre eux pouvant être continu, ordinal ou discret. Ensuite, même dans le cas d'un seul facteur à deux niveaux, notre approche ne nécessite pas de classification ou d'appariement d'après le score de propension, de telle sorte que le risque de "confusion résiduelle" (c'est-à-dire de biais) lié à ces méthodes est évité. Enfin, notre approche permet de conforter l'idée qu'il vaut mieux utiliser le score de propension estimé que le vrai score de propension, même lorsque ce vrai score est connu. Le surcroît de puissance de notre approche provient du fait que nous supposons que l'influence des facteurs peut être décrite par la composante paramétrique d'un modèle de régression semi-paramétrique. A titre d'illustration, nous réanalysons, sur une cohorte de 2,713 adultes blancs de sexe masculin, l'effet du tabac sur la valeur du volume expiratoire maximal seconde, et nous comparons les résultats obtenus avec ceux des méthodes classiques.

References

Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* 11, 432–452.

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1992). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–334.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dockery, D. W., Spelzer, F. E., Ferris, B. G., Ware, J. H., Louis, T. A., and Spiro, A. (1988). Cumulative and reversible effects of lifetime smoking on simple tests of lung function in adults. *American Review of Respiratory Diseases* 137, 286–292.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* 65, 657–687.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71, 431–444.

Huber, P. (1981). *Robust Statistics*. New York: Wiley.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* 79, 61–71.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear model. *Biometrika* 73, 13–22.

Manski, C. F. (1988). *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* 10, 475–478.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7, 1393–1512.

Robins, J. M. (1989a). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman, and A. Mulley (eds), 113–159. Washington, DC.: NCHSR, U.S. Public Health Service.

Robins, J. M. (1989b). The control of confounding by intermediate variables. *Statistics in Medicine* 8, 679–701.

Robins, J. M. (1992a). Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics*, in press.

Robins, J. M. (1992b). Estimating the causal effect of a time-varying treatment on survival using a new class of failure time models. *Communications in Statistics*, in press.

Robins, J. M. (1992c). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, in press.

Robins, J. M. (1992d). Analytic methods for HIV treatment and cofactor effects. In *Methodological Issues of AIDS Behavioral Research*, D. G. Ostrow and R. Kessler (eds). New York: Plenum.

produced by RJRI

522299 4757

13 of 17

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Generated from www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:46 PM FAX NO. P. 19

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia (PCP) on the survival of AIDS patients. *Epidemiology*, in press.

Robins, J. M. and Morgenstern, H. (1987). The foundations of confounding in epidemiology. *Computers and Mathematics with Applications* 14, 869–916.

Robinson, P. (1988). Root-*N*-consistent semiparametric regression. *Econometrica* 56, 931–954.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* 79, 565–574.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* 82, 387–394.

Rosenbaum, P. R. (1988). Permutation tests for matched pairs with adjustments for covariates. *Applied Statistics* 37, 401–411.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 33–38.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6, 34–58.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.

Received December 1989; *revised October* 1990 *and January* 1991; *accepted February* 1991.

Appendix

In this Appendix, we prove the results stated in the text. We assume that

$$Y_i = f(S_i, X_i, \beta) + h(X_i) + \varepsilon_i, \quad E[\varepsilon_i | S_i, X_i] = 0, \tag{A.0}$$

where $f(S_i, X_i, \beta)$ is a linear function of a V-dimensional parameter vector β that takes the value zero when $S_i = 0$. [Extension of our results to nonlinear functions of β is straightforward.] Model (2) in the text is the special case in which $f(S_i, X_i, \beta) = \beta S_i$ for univariate β and dichotomous S_i. (A.0) generalizes (2) by allowing for multivariate exposures, each component of which may be categorical, ordinal, or continuous. For example, we might suppose $S_i = (S_{i1}, \ldots, S_{iM})$ and $f(S_i, X_i, \beta) = \sum_{m=1}^{M-1} \beta_m S_{im} + \beta_{M+1} S_{iM} X_i$ with $V = M + 1$. If equation (4b) holds when s is any value of S_i, then $f(S_i, X_i, \beta)$ is the average effect of joint exposure level S_i compared to the baseline level $S_i = 0$ among subjects with covariate level X_i. If $f(S_i, X_i, \beta)$ depends on X_i, we say there is an exposure-covariate interaction.

Define $f_\beta(S_i, X_i)$ to be the V-vector of partial derivatives of $f(S_i, X_i, \beta)$ with respect to the components of β and let

$$E[f_\beta(S_i, X_i) | X_i] = r(X_i; \alpha), \tag{A.1}$$

where $r(\cdot; \cdot)$ is a known function and α is an unknown parameter. Define $R(S_i, X_i; \alpha) \equiv f_\beta(S_i, X_i) - r(X_i; \alpha)$. Note that $E[R(S_i, X_i; \alpha) | X_i] = 0$. If, as in the text, $f_\beta(S_i, X_i) = S_i$ is a Bernoulli random variable, (A.1) is a fully parametric model for S_i given X_i; otherwise, (A.1) is a semiparametric model for the density $f_\beta(S_i | X_i)$, since the distribution of $R(S_i, X_i; \alpha)$ is completely unrestricted except for having mean zero given X_i.

Now for any nonrandom function $g(x)$, define

$$\begin{aligned} n^{-1/2} U(\beta^*, g, \hat{\alpha}) &= n^{-1/2} \sum_i [Y_i - f(S_i, X_i, \beta^*) - g(X_i)] R(S_i, X_i; \hat{\alpha}) \\ &= n^{-1/2} \sum_i U_i(\beta^*, g, \hat{\alpha}), \end{aligned} \tag{A.2}$$

where $\hat{\alpha}$ is asymptotically equivalent to an $n^{1/2}$-consistent solution to $0 = \sum_i M_i(\alpha^*) = \sum_i m(S_i, X_i, \alpha^*)$ for some $M_i(\alpha^*)$ satisfying $E[M_i(\alpha)] = 0$. That is, when $M_i(\alpha)$ is continuously differentiable, $n^{1/2}(\hat{\alpha} - \alpha) = -[E[\partial M_i(\alpha)/\partial \alpha']]^{-1} n^{-1/2} \sum_i M_i(\alpha) + o_p(1)$, and we say that $-[E[\partial M_i(\alpha)/\partial \alpha']]^{-1} M_i(\alpha)$ is the influence function of $\hat{\alpha}$. Chamberlain (1987) proves that $\hat{\alpha}$ is semiparametric efficient for α

52299 4758

14 of 17

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf Powered by http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-88 THU 08:47 PM FAX NO. P. 20

under the sole restriction (A.1) on the conditional distribution of S_i given X_i only if $M_i(\alpha)$ equals

$$M_i^{\pi}(\alpha) \propto \frac{\partial r(X_i; \alpha)}{\partial \alpha} [\mathrm{var}[R(S_i, X_i; \alpha)|X_i]]^{-1} R(S_i, X_i; \alpha).$$

That is, $\hat{\alpha}$ is semiparametric efficient only if it is asymptotically equivalent to the *optimal* weighted (possibly nonlinear) least squares estimator of α. Henceforth, we shall say that $\hat{\alpha}$ is semiparametric efficient under (A.1) if $\hat{\alpha}$ has influence function $-[E[\partial M_i^{\pi}(\alpha)/\partial \alpha']]^{-1} M_i^{\pi}(\alpha)$. If $f_\theta(S_i, X_i) \equiv S_i$ is Bernoulli, semiparametric efficiency under (A.1) is just ordinary parametric efficiency. Our main result is given as Theorem A.1.

Theorem A.1 Under regularity conditions given in Corollary 1, Chapter 8 of Manski (1988), there exists a solution $\hat{\beta}_L(g) \equiv \hat{\beta}_L(g, \hat{\alpha})$ to $n^{-1/2} U(\beta^*, g, \hat{\alpha}) = 0$ such that $n^{1/2}(\hat{\beta}_L(g) - \beta)$ is asymptotically normal with mean 0 and variance that can be consistently estimated by

$$\hat{J}^{-1} \hat{\Delta}(g) (\hat{J}')^{-1}, \tag{A.3a}$$

where

$$\hat{J}' \equiv n^{-1} \sum_i \partial \hat{U}_i(\beta, g, \hat{\alpha})'/\partial \beta = n^{-1} \sum_i f_\theta(S_i, X_i) R(S_i, X_i; \hat{\alpha})',$$

$$\hat{\Delta}(g) = n^{-1} \sum_i K_i(\hat{\beta}, g, \hat{\alpha}) K_i'(\hat{\beta}, g, \hat{\alpha}),$$

$$K_i(\hat{\beta}, g, \hat{\alpha}) = U_i(\hat{\beta}, g, \hat{\alpha}) - \hat{B}(g) \hat{C}^{-1} M_i(\hat{\alpha}),$$

$$\hat{\beta} \equiv \hat{\beta}_L(g),$$

$$\hat{B}(g) \equiv n^{-1} \sum_i \partial U_i(\hat{\beta}, g, \hat{\alpha})/\partial \alpha' = n^{-1} \sum_i [Y_i - f(S_i, X_i, \hat{\beta}) - g(X_i)] \frac{\partial R(S_i, X_i; \hat{\alpha})}{\partial \alpha'},$$

$$\hat{C} \equiv n^{-1} \sum_i \partial M_i(\hat{\alpha})/\partial \alpha'.$$

If $\hat{\alpha}$ is semiparametric efficient under (A.1), the asymptotic variance of $n^{1/2}(\hat{\beta}_L(g) - \beta)$ can be consistently estimated by

$$\hat{J}^{-1} \hat{\Sigma}(g) (\hat{J}')^{-1} = \hat{Q}(g) \hat{\Omega} \hat{Q}'(g), \tag{A.3b}$$

where $\hat{Q}(g) \equiv \hat{J}^{-1} \hat{B}(g)$, $\hat{\Sigma}(g) = n^{-1} \sum_i U_i(\hat{\beta}_L(g), g, \hat{\alpha}) U_i(\hat{\beta}_L(g), g, \hat{\alpha})'$, and $\hat{\Omega}$ is a consistent estimator of $\mathrm{var}^{\infty}[n^{1/2}(\hat{\alpha} - \alpha_n)]$.

Except when $f_\theta(S_i, X_i)$ equals a dichotomous S_i (as in the text), $\mathrm{var}[R(S_i, X_i; \alpha)|X_i]$ may be an unknown function of α and X_i. Hence, if one chooses to estimate α by the unweighted (possibly nonlinear) least squares regression of $f_\theta(S_i, X_i)$ on X_i, it is necessary to use formula (A.3a) rather than (A.3b), since $\hat{\alpha}$ will then be efficient only if the (unknown) variance of $R(S_i, X_i; \alpha)$ does not depend on X_i.

However, if one has a correctly specified model $\mathrm{var}[R(S_i, X_i; \alpha)|X_i] = \psi(X_i; \theta)$, where $\psi(X_i; \theta')$ is a known function and θ is an unknown parameter, then it is well known that the estimate $\hat{\alpha}$ that solves $0 = \sum_i [\partial r(X_i; \alpha)/\partial \alpha][\psi(X_i; \hat{\theta})]^{-1} R(S_i, X_i; \alpha)$ has influence function $-[E[\partial M_i^{\pi}(\alpha)/\partial \alpha']]^{-1} M_i^{\pi}(\alpha)$ and (A.3b) can be used. Here $\hat{\theta}$ is the (possibly nonlinear) multivariate least squares regression estimate of θ obtained by regressing $R(S_i, X_i; \hat{\alpha}) R'(S_i, X_i; \hat{\alpha})$ on X_i, where $\hat{\alpha}$ is obtained from a preliminary unweighted least squares regression of $f_\theta(S_i, X_i)$ on X_i.

Application of Theorem A.1 Consider equation (9) in the text. In that setting, $g(X_i) \equiv 0$; $f(S_i, X_i, \beta) = \beta S_i$; $f_\theta(S_i, X_i) = S_i$; $R(S_i, X_i; \hat{\alpha}) \equiv S_i - \hat{p}_i$, where $\hat{p}_i \equiv e^{\hat{\alpha}' X_i}/(1 + e^{\hat{\alpha}' X_i})$;

$$\frac{\partial R(S_i, X_i; \hat{\alpha})}{\partial \alpha'} = \frac{\partial}{\partial (\alpha')} [e^{(\alpha_n)' X_i}/(1 + e^{(\alpha_n)' X_i})]|_{\alpha' = \alpha_n} = \hat{p}_i(1 - \hat{p}_i) X_i';$$

$Y_i - f(S_i, X_i, \hat{\beta}_L(g)) - g(X_i) \equiv \hat{z}_i$; $U_i(\hat{\beta}_L(g), g, \hat{\alpha}) = \hat{z}_i(S_i - \hat{p}_i)$; $\hat{J} = n^{-1} \sum_i S_i(S_i - \hat{p}_i)$; $\hat{\Sigma}(g) \equiv n^{-1} \sum_i \hat{z}_i^2 (S_i - \hat{p}_i)^2$; $\hat{Q}(g) \equiv \sum_i \hat{z}_i \hat{p}_i (1 - \hat{p}_i) X_i' / \sum_i S_i (S_i - \hat{p}_i)$. Substituting into equation (A.3b), we obtain equation (9).

The reader can check that substituting in (A.3b) also gives equation (13) if we set $g(X_i) \equiv \hat{\beta}_1 + \sum_{k=2}^{K} \hat{\beta}_k X_{ki}$ above. [As we shall see in Theorem (A.3) below, the fact that $g(X_i)$ is based on estimates $\hat{\beta}_k$ does not affect the asymptotic variance (A.3b).]

52299 4759

http://legacy.library.ucsf.edu/tid/ocq07a00/pdf Powered by UCSF Industry Documents Library http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:48 PM FAX NO. P. 21

Proof of Theorem A.1 For pedagogic purposes we first sketch a proof. We then show how Corollary 1 of Manski's Chapter 8 can be used to formally prove the theorem. Since

$$E[U_i(\beta, g, \alpha)] = 0 \tag{A.4}$$

by (A.0), we have that, under the regularity conditions discussed below, a Taylor expansion and the weak law of large numbers (WLLN) gives

$$0 = n^{-1/2}U(\hat{\beta}_E(g), g, \hat{\alpha}) = n^{-1/2}U(g) + I[n^{1/2}(\hat{\beta}_E(g) - \beta)] + B(g)[n^{1/2}(\hat{\alpha} - \alpha)] + o_p(1),$$

where $n^{-1/2}U(g) \equiv n^{-1/2}\Sigma U_i(g) \equiv n^{-1/2}U(\beta, g, \alpha)$, $I \equiv E[\partial U_i(\beta, g, \alpha)/\partial\beta']$, which does not depend on g.

$$B(g) \equiv E\left(\frac{\partial U_i(\beta, g, \alpha)}{\partial \alpha'}\right).$$

Thus

$$n^{1/2}(\hat{\beta}_E(g) - \beta) = -I^{-1}[B(g)n^{1/2}(\hat{\alpha} - \alpha) + n^{-1/2}U(g)] + o_p(1). \tag{A.5}$$

By assumption, $n^{1/2}(\hat{\alpha} - \alpha) = -C^{-1}n^{-1/2}\Sigma_i M_i + o_p(1)$, where $C \equiv E[\partial M_i(\alpha)/\partial\alpha]$, $M_i \equiv M_i(\alpha)$. Hence $n^{1/2}(\hat{\beta}_E(g) - \beta) = -I^{-1}n^{-1/2}\Sigma_i[U_i(g) - B(g)CM_i] + o_p(1)$. Thus $n^{1/2}(\hat{\beta}_E(g) - \beta)$ is asymptotically normal with mean zero and variance $I^{-1}\Delta(g)(I')^{-1}$, where $\Delta(g) \equiv \text{var}[U_i(g) - B(g)C^{-1}M_i]$ since $n^{1/2}(\hat{\beta}_E(g) - \beta)$ is a sum of independent mean-zero random variables plus a term of $o_p(1)$. Formula (A.3a) follows by the WLLN.

We next establish (A.3b) for a semiparametric efficient $\hat{\alpha}$ under (A.1) using arguments similar to those in Pierce (1982) and Newey (1990). Let $L_i(\alpha^\dagger, \eta^\dagger) \equiv f(S_i|X_i; \alpha^\dagger, \eta^\dagger)$ be any (regular) parametric submodel with true values α, η for the density of S_i given X_i consistent with the restriction (A.1). Let $S_{i\alpha} \equiv \partial \ln L_i(\alpha, \eta)/\partial\alpha^\dagger$. Let

$$\tau \equiv \{a(S_i, X_i): a(S_i, X_i) \equiv \partial \ln L_i(\alpha, \eta)/\partial\eta^\dagger \text{ for some parametric submodel}\}.$$

Note $\tau = \{a(S_i, X_i): E[a(S_i, X_i)|X_i] = 0$ and $E[R(S_i, X_i; \alpha)a(S_i, X_i)'|X_i] = 0\}$ since the scores $a(S_i, X_i)$ are restricted only by having a conditional mean of zero and by being conditionally uncorrelated with $R(S_i, X_i; \alpha)$. It follows from Chamberlain (1987), Begun et al. (1983), and Newey (1990) that (a) $S_{i\alpha} - M_i^{*\tau} \in \tau$ and $E[M_i^{*\tau}a(S_i, X_i)'] = 0$ for all $a(S_i, X_i) \in \tau$ and (b) $\text{var}^A[n^{1/2}(\hat{\alpha} - \alpha)] = [E(M_i^{*\tau}(M_i^{*\tau})')]^{-1}$. $M_i^{*\tau}$ is called the efficient score in the semiparametric model (A.1) for the law of S_i given X_i.

Now by differentiating the identity $E_{\alpha,\eta,\alpha^\dagger}[U_i(\beta, g, \alpha^\dagger)] = 0$ with respect to $\alpha^\dagger$ using the chain rule and evaluating at the true values (α, η), we obtain $B(g) = -E[U_i(g)S'_{i\alpha}]$, where $E_{\alpha,\eta,\alpha^\dagger}$ refers to expectation with respect to a density that differs from the truth only in that the law of S_i given X_i is $f(S_i|X_i; \alpha^\dagger, \eta^\dagger)$. Similarly differentiating this identity with respect to $\eta^\dagger$, we obtain $E[U_i(g)a'(S_i, X_i)] = 0$ for all $a(S_i, X_i) \in \tau$. Thus, by (a) in the last paragraph, we conclude $B(g) = -E[U_i(g)(M_i^{*\tau})']$. Similarly, the identity $E_{\alpha,\eta,\alpha^\dagger}[M_i(\alpha^\dagger)] = 0$ implies $C = -E[M_i(M_i^{*\tau})']$. Hence $K_i(g) \equiv U_i(g) - B(g)C^{-1}M_i = U_i(g) - E[U_i(g)(M_i^{*\tau})']\{E[M_i(M_i^{*\tau})']\}^{-1}M_i$. In the special case in which $M_i = M_i^{*\tau}$, $K_i(g)$ is the residual from the (population) least squares regression of $U_i(g)$ on $M_i^{*\tau}$, and a standard calculation gives $\text{var}[K_i(g)] = \text{var}[U_i(g)] - B(g)C^{-1}B'(g)$. (A3.b) then follows by (b) in the last paragraph and the WLLN.

Theorem A.1 is formally proved by noting that it is an immediate consequence of Corollary 1 in Manski's Chapter 8 and the above variance calculations when we set Manski's function $g(z, b)$ equal to $(U_i(\beta', g, \alpha^\dagger)', M_i(\alpha^\dagger)')'$ and Manski's function $r(x)$ equal to $x'x$, where x is a vector.

Corollary A.1 If $\hat{\alpha}^{(j)}$ is semiparametric efficient under the jth of J nested correctly specified models $E[f_j(S_i, X_i)|X_i] = r(X_i; \alpha^{(j)})$, $(j = 1, \ldots, J)$, with the dimension of $\alpha^{(j)}$ increasing with j, then the asymptotic variance of $\hat{\beta}_E{}^{(j)}(g) \equiv \hat{\beta}_E(g, \hat{\alpha}^{(j)})$ is nonincreasing with j.

Proof Correct specification implies that, for $j > j^*$, $M_i^{*\tau^j}$ is the first j^* components of $M_i^{*\tau^j}$, the efficient score for the jth model. But, by standard least squares theory, the variance of the residual $K_i^{(j)}(g)$ based on the jth model must be less than or equal to that based on model j^*.

The following theorem will be used in proving the claims made in the paragraph following equation (13).

Theorem A.2

(a) $\text{var}^A[n^{1/2}(\hat{\beta}_E(g) - \beta)] \geq \text{var}^A[n^{1/2}(\hat{\beta}_E(h) - \beta)]$.
(b) $\text{var}^A[n^{1/2}(\hat{\beta}_E(h) - \beta)] = \text{var}^A[n^{1/2}(\hat{\beta}_L(h) - \beta)]$, where $\hat{\beta}_L(h) \equiv \hat{\beta}_L(h, \alpha)$ and $\hat{\beta}_E(h) \equiv \hat{\beta}_E(h, \hat{\alpha})$.

Produced by R.J.R.

52299 4760

16 of 17

http://legacy.library.ucsf.edu/tid/ooq07a00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001

JUL-01-99 THU 08:49 PM FAX NO. P. 22

Proof of (a) (a) is an immediate consequence of the following two lemmas.

Lemma A.1 The function g minimizing $\text{var}^\wedge[n^{1/2}(\hat{\beta}_E(g) - \beta)]$ also minimizes $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})]$.

Proof By a Taylor expansion, we have

$$0 = n^{-1/2}U(\hat{\beta}_L(g), g, \hat{\alpha}) = n^{-1/2}U(\beta, g, \hat{\alpha}) + n^{-1}\frac{\partial U(\beta, g, \hat{\alpha})}{\partial \beta'}[n^{1/2}(\hat{\beta}_E(g) - \beta)] + o_p(1). \quad (A.6)$$

A further Taylor expansion of $n^{-1}\partial U(\beta, g, \hat{\alpha})/\partial\beta'$ around α_0 and the WLLN proves $n^{-1}\partial U(\beta, g, \hat{\alpha})/\partial\beta' = I + o_p(1)$, proving the lemma.

Lemma A.2 The function h minimizes $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})]$.

Proof $n^{-1/2}U(\beta, g, \hat{\alpha}) = n^{-1/2}\sum_i \varepsilon_i R(S_i, X_i; \hat{\alpha}) + n^{-1/2}\sum_i [h(X_i) - g(X_i)]R(S_i, X_i; \hat{\alpha}) \equiv A_1 + A_2(g)$, say where we have used (A.0) to substitute $\varepsilon_i + h(X_i)$ for $Y_i - f(S_i, X_i, \beta)$. If we can show $\text{cov}^\wedge(A_1, A_2(g)) = 0$, then $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})] = \text{var}^\wedge(A_1) + \text{var}^\wedge[A_2(g)]$, which is minimized at $g = h$ since $\text{var}^\wedge[A_2(h)] = 0$. Now A_1 and $A_2(g)$ have zero covariance since (a) $E[A_1|(S, X)] = 0$ and (b) $A_2(g)$ is fixed given $(S, X) \equiv [(S_i, X_i); i = 1, \ldots, n]$. (a) and (b) follow from the fact that $E[\varepsilon_i|(S, X)] = 0$ and $\hat{\alpha}$ depends on the data only through (S, X).

Proof of (b) (b) follows from the fact that $B(h) = 0$ by (A.0).

In general, we do not know $h(X_i)$. Therefore, as in Section 4, we shall hypothesize a model $h(X_i) = g(X_i, \theta)$ where $g(\cdot, \cdot)$ is a known function and θ is a vector of parameters to be estimated. We estimate θ by (possibly nonlinear) least squares regression of $Y_i - f(S_i, X_i, \hat{\beta}_E)$ on X_i, where $\hat{\beta}_E$ is $\hat{\beta}_E(g)$ for $g(X) \equiv 0$. Let $\tilde{\theta}$ be the (possibly nonlinear) least squares estimator of θ. It is clear that, since $\hat{\beta}_E$ is an $n^{1/2}$-consistent estimator of β, if the model for $h(X_i)$ were correctly specified, $n^{1/2}(\tilde{\theta} - \theta)$ would have a nondegenerate limiting distribution with mean 0. If the model for $h(X_i)$ were misspecified, there still exists θ^* such that $n^{1/2}(\tilde{\theta} - \theta^*)$ has a nondegenerate limiting distribution with mean 0. The following theorem shows that we can then use $\tilde{\theta}$ to construct an adaptive estimator of β that (1) has the same limiting distribution as $\hat{\beta}_E(h)$ if our model $h(X_i)$ is correctly specified and (2) remains consistent, asymptotically normal even if our model is misspecified.

Theorem A.3 If $n^{1/2}(\tilde{\theta} - \theta^*)$ has a nondegenerate limiting distribution with mean 0, then $\hat{\beta}_E[g(X_i, \tilde{\theta})]$ has the same limiting distribution as $\hat{\beta}_E[g(X_i, \theta^*)]$. In particular, it will be consistent and asymptotically normal whether or not the hypothesized model for $h(X_i)$ is correct, and it will have the same limiting distribution as $\hat{\beta}_E(h)$ if the model for $h(X_i)$ is correct.

Proof For notational convenience, assume that θ is one-dimensional. It will be sufficient to show that

$$n^{-1/2}U(\beta', \hat{\alpha}, g(\tilde{\theta})) = n^{-1/2}(U(\beta', \hat{\alpha}, g(\theta^*)) + o_p(1) \quad (A.7)$$

for $|\beta' - \beta| = O(n^{-1/2})$. By a Taylor series expansion

$$n^{-1/2}U(\beta', \hat{\alpha}, g(\tilde{\theta})) = n^{-1/2}(U(\beta', \hat{\alpha}, g(\theta^*)) + n^{1/2}(\tilde{\theta} - \theta^*)[n^{-1}U[\beta', \hat{\alpha}, g'(\theta^*)] \quad (A.8)$$

$$+ n^{1/2}(\tilde{\theta} - \theta^*)^2[n^{-1}U[\beta', \hat{\alpha}, g''(\tilde{\theta}^*)]] \quad (A.9)$$

for some $\tilde{\theta}^*$ between $\tilde{\theta}$ and θ^*. Now, if $\beta' = \beta$, by Theorem A.1 and Pierce (1982), $[n^{-1}U[\beta', \hat{\alpha}, g'(\theta^*)]]$ converges to 0 in probability since it has mean 0 to $o_p(n^{-1/2})$ with variance converging to 0 as $n \to \infty$. Further, under regularity conditions, this remains true if $|\beta' - \beta| = O(n^{-1/2})$. It then follows from Slutsky's theorem that expression (A.8) converges in law to 0 and thus in probability to 0. Further, since $n^{-1}U[\beta', \hat{\alpha}, g''(\tilde{\theta}^*)]$ is at most $O_p(1)$ and $n^{1/2}(\tilde{\theta} - \theta^*)^2$ is $O_p(n^{-1/2})$, it follows that expression (A.9) is $O_p(n^{-1/2})$. Thus equation (A.7) is true.

Theorem (A.3) and part (b) of Theorem (A.2) imply proposition (1) in the paragraph following equation (13). Proposition (2) is an easy calculation.

52299 4761

17 of 17

http://legacy.library.ucsf.edu/tid/oeq07a00/pdf
Stanford Libraries http://www.industrydocuments.ucsf.edu/docs/hygl0001